### Membrane Permeability Database for Nonpeptidic Macrocycles

Qiushi Feng1, Danjo De Chavez1, Roland Lindh1, Jan Kihlberg1, Vasanthanathan, Poongavanam1

**Affiliations**

1. Department of Chemistry-BMC, Uppsala University, SE-75123, Uppsala, Sweden

Corresponding author(s): Jan Kihlberg ([jan.kihlberg@kemi.uu.se](mailto:jan.kihlberg@kemi.uu.se)); Vasanthanathan Poongavanam ([Vasanthanathan.poongavanam@kemi.uu.se](mailto:Vasanthanathan.poongavanam@kemi.uu.se))

### Abstract

The process of developing new drugs from discovery to market is arduous and costly, particularly for targets classified as "difficult-to-drug." Traditional molecules adhering to Lipinski's Rule of Five (Ro5) often fail to address these challenging targets. Biologics, while suitable for some, face limitations in cell permeability, hindering their efficacy for intracellular targets and oral administration. Molecules beyond Ro5 boundaries, notably macrocycles, have shown promise in modulating difficult-to-drug targets, including Protein-Protein Interactions (PPIs), while still permitting oral administration. However, assessing cell permeability, critical for intracellular targeting and oral bioavailability, is laborious and expensive through experimental methods. In silico methods offer a cost-effective and accurate alternative, enabling predictions prior to compound synthesis. However, due to the limited number of non-peptidic macrocycles with experimental data, hindering prediction. Here, we present a comprehensive online database (http://swemacrocycle.uu.se), housing cell permeability data for nonpeptidic macrocycles curated from diverse literature, patents, and bioactivity repositories. The database contains 4216 unique macrocycles, facilitating accurate computational predictions for drug discovery projects. This resource fills a critical gap in existing databases by including nonpeptidic macrocycles, offering valuable insights into their membrane permeability prediction, and facilitating the development of novel therapeutics.

**Graphical abstract**

A diagram of data processing

Description automatically generated

### Background & Summary

Developing a new drug from discovery to market is an expensive and time-consuming process.1 Approximately half of the targets associated with human diseases are classified as "difficult-to-drug" with traditional molecules following Lipinski's Rule of Five (Ro5)2, which outlines limits for molecular weight (MW <500 Da), calculated lipophilicity (cLogP <5), as well as hydrogen bond donors and acceptors (HBD <5, HBA <10). Although biologics may be suitable for challenging targets, their lack of cell permeability hinders access to intracellular targets and renders them unsuitable for oral administration. Ongoing research has shed light on compounds reside outside the Ro5 boundaries, known as beyond Rule of Five (bRo5 space).3-5 Among these compounds, macrocycles, characterized by a ring of at least 12 heavy atoms, exhibit the capability to modulate difficult-to-drug targets, including those with extensive, flat, or groove-shaped binding sites, as well as Protein-Protein Interactions (PPIs), while still allowing for oral administration (e.g., cyclosporin A).6

Measurement of cell permeability of drugs is not only crucial to assess their ability to reaching intracellular targets, regardless of their location in the central nervous system (CNS) or peripheral sites,7 also utilized as a model system for estimating intestinal absorption. Various *in vitro* assays are employed to measure cell permeability and these assays include the human colorectal carcinoma cell (Caco-2), the parallel artificial membrane permeability assay (PAMPA), Madin–Darby canine kidney (MDCK) cells, and the low-efflux MDCK clone Ralph Russ canine kidney (RRCK), etc. However, generating experimental permeability data is both time-consuming and expensive. Alternatively, *in silico* methods are not only cost-effective but also sufficiently accurate and fast enough to be used as high-throughput filter in the drug discovery projects, enabling predictions before compound synthesis and testing.8-10

Despite the emergence of various cyclic peptides as a promising chemical class in drug discovery,11-13 offering improved design strategies and cost-effective synthesis, compared to nonpeptidic macrocycles (NPM), they often encounter issues of metabolic instability. In contrast, nonpeptidic macrocycles provide a more favourable chemical space for modulating *difficult-to-drug* targets, despite their poor membrane permeability—a critical factor in assessing oral bioavailability and intracellular targeting potential. The challenges associated with poor membrane permeability are further exacerbated by limited experimental data on how macrocycles traverse cellular membranes.

To facilitate the development of accurate and efficient computational predictions, it is crucial to collect and curate experimental data with structural information, making it available to scientific communities. Our work is inspired by the recent work of Li J et al.,14 CycPeptMPDB, a comprehensive database of membrane permeability for cyclic peptides containing more than 7000 cyclic peptides from several publications. However, Li J work is limited to cyclic peptide and do not include any NPM. In this study, we have developed a cell permeability database for nonpeptidic macrocycles—a comprehensive online resource for macrocycles with cell permeability. Here, we have collected and curated macrocycles (both semi-peptide and non-peptide) from scientific literature, patents, and various bioactivity data repositories. The database comprises structures and cell permeability data obtained from different assays, endpoints, and molecular features for 4602 unique macrocycles, which are easily accessible and downloadable through the web server (<http://swemacrocycle.com>) for further modeling purpose. The overall workflow of the present study is shown in **Fig 1**.

A diagram of data processing

Description automatically generated

**Figure 1:** The workflow of the cell permeability database for nonpeptidic macrocycles. Structures and permeability data were retrieved from the literature, patents, and scientific databases, followed by manual curation. The webserver provides readily downloadable datasets for various cell permeability assays and endpoints. Additionally, the webserver incorporates analyses based on various physicochemical descriptors.

### Methods

**Data collection and curation**

Macrocycles exhibiting cell permeability were obtained from three different sources: 1) scientific literature, 2) patents, and 3) public repositories, as detailed below. In the scientific literature, a search was conducted on PubMed15 and Google Scholar to identify published articles reporting information on macrocycles with cell permeability. Keywords like “macrocycle” were combined with either the general term “cell permeability” or specific assay names (Caco-2, PAMPA, MDCK, RRCK) to query the top medicinal chemistry journals. Similar search was done on Google Patents. More permeability data were retrieved and collected from ChEMBL by using its Python web resource client (Ref. Nucleic Acids Research, 2015, Vol. 43), through querying the same keywords mentioned before (Python script can be found in Supporting information) .

The RDKit Molecule Substructure module17 was employed to filter macrocycles, defined as a ring with at least 12 heavy atoms. Subsequently, all SMILES and cell permeability data were imported into Molecular Operating Environment (MOE version xxx).18 The dataset underwent curation, involving the removal of mixtures, inorganics, salts, solvent molecules, and structural normalization. This resulted in more than 4500 diverse and unique macrocycles.

**Quantification of cyclic and non-cyclic peptides**

Given the absence of a standardized quantitative definition for macrocycles encompassing both cyclic and non-cyclic peptides, we propose the calculation of the ratio between the total number of backbone amide atoms and the size of the macrocyclic ring. This computation aims to incorporate both amide bonds that constitute the peptide linkage within macrocycles, thereby providing insight into the peptide nature of the macrocycle.

AR = (3 × FAB) / MRS

Where, FAB denotes the total number of amide bonds (inclusive of NH and N-alkylated bonds), each multiplied by three to account for the number atoms (-N-Ca-C-) forming each amide bond within the macrocyclic ring. MRS represents the total number of atoms in the macrocyclic ring. AR returns a value between 0 and 1, a value of 0 represents a non-peptide macrocycle, and 1 represent cyclic peptide. CycPeptMPDB14 represented as a peptide dataset. Non-peptide macrocycles were collected from various publications 6,8,19. Publications that specially refer themselves as semi-peptides were included as semi-peptide class.20-22

**Webserver implementation**

The web server implemented in this study was built upon the Django web framework (version 3.2.23). The development of the web interface involved the use of standard web technologies, including HTML5, CSS, and JavaScript, with all data within the web server stored and managed using SQLite, a lightweight and efficient relational database management system. For molecule visualization, RDKit (version 2023.9.5) 23 was employed. Specifically, RDKit was used to generate 2D images of molecules24 and convert SMILES into PNG image and SDF files within the web interface. To support online data visualization, ECharts (version v5.5.0) was utilized. 25 The functionality for table sorting and filtering was implemented using DataTables (https://www.datatables.net/), a JavaScript library for enhancing HTML tables. The website has been thoroughly tested to ensure functionality across multiple operating systems and web browsers. Most of the codes used in this work are open-source and properly acknowledged. The code for the final version of the web server is also provided on GitHub.

### Data Records

All the macrocycles, along with their cell permeability data and molecular features, are available for the user on the <http://swemacrocycle.com>web server. For each data point, a unique molecule ID is provided, which may correspond to multiple cell permeability assay measurements reported. We have divided the data record for each entry into three categories: (i) *representation*, where we provide 2D structure, Isomeric SMILES, InChI Key for the overall structure and macrocyclic ring core; (ii) *Cell permeability* information containing the type of cell permeability assay, endpoints, values, and units. If there are more than one assay reported for the same compound, we provide all of them in a separate entry. Since each molecule carries a unique entry, it is easier for the user to check the different assays and values; (iii) Each entry comes with key *molecular features*, including Lipinski26 and Veber27 descriptors. In addition, amide ratio and the fraction of sp3 carbon atoms are provided for each entry.

Macrocycles possess a vast conformational space and are challenging to predict relevant conformations with molecular mechanics, irrespective of solvation models and force fields used in the conformational sampling procedure.28 However, the QM-based conformational sampling procedure finds relevant conformations of macrocycles29, due to the high computational cost in sampling macrocycles provided in the present work. We provide Isomeric SMILES, cell permeability, and key 2D molecular features. In addition to the above information for each entry, the original sources from which the structure and cell permeabilities were extracted are also provided for the user. For ready-to-use dataset for QSAR modeling, all the cell permeability values were standardized into logarithmic values. Moreover, all data records incorporated in the web server are ready to download with multiple search and sorting options (**Fig. 2**).

A screenshot of a cell permeability chart

Description automatically generated

**Fig. 2.** Schematic representation of macrocyclic report for each entry in the database is shown.

*Abbreviations*: MW: Molecular Weight; HBA: Hydrogen bond acceptor; HBD: Hydrogen bond donor; cLogP: Calculated lipophilicity; TPSA: Topological polar surface area; NRotB: Number of rotatable bonds; Φ: Kier flexibility Index: Fsp3: fraction of sp3 carbon atoms; MRS: macrocyclic ring size.

### Technical Validation

**Cell permeability database statistics. *Source:*** Macrocycles reported here were obtained from both journals, and patents, consists of 5646 macrocycles collected from 103 journal publications and 9 patents during 2006-2023 (*last updated July 2023*). Out of 5646 macrocycle records, 84%, 4%, and 11% of macrocycles are from journal articles, patents, and the ChEMBL repository, respectively. ***Assay:*** In total, the compounds provided in the database are divided into five major cell permeability assay categories (**Fig. 3b**), namely, PAMPA, Caco-2, MDCK, RRCK, and others (or none of the above). Macrocycles with PAMPA-based passive permeability records account for 67%, among which 91% (3462 macrocycles) are from one publication.30 This publication contains log (Peff) measured under consistent experimental conditions, making it one of the largest sources of macrocyclic data with cell permeability available in the public domain. The next highest number of data records is based on the Caco-2 assay, comprising 26% (1504) of macrocycles. Approximately 450 compounds had efflux ratio (ER) data, a measure of the ratio of Papp in the apical to basal (AB) direction to the basal to apical (BA) direction, often used to determine whether a compound undergoes active efflux. Moreover, efflux ratio with inhibitor (ER+) is also provided for 111 macrocycles. Out of 1504 macrocycles, log (Papp) AB and log (Papp) BA data were available for 415 and 337 macrocycles, respectively, with and without cocktail efflux inhibitors. Another commonly used cell-based permeability assay, MDCK, has 268 compounds with cell permeability data, including 11 compounds with efflux ratio and 118 compounds with log (Papp) AB. Not many macrocycles have RRCK data (*n=7*), and there are also 98 compounds that have cell permeability data but do not belong to any of the above assay categories. However, log (Papp) AB is provided for 52 compounds and ER for 46 compounds, for which the cell permeability is primarily measured with LLC-PK1 cells.

A screenshot of a computer

Description automatically generated

**Figure 3:** a) Nested-pie chart of cell permeability data for macrocycles. b)TMAP visualization of the non-peptidic macrocyclic dataset (*n=5646*). Various assays and features are highlighted. (c) Molecular property distribution of macrocycles included in the database is shown, with 2D molecular features representing size, lipophilicity, polarity, flexibility, and amide ratio.

**Macrocycle diversity:** Structural diversity and permeability data were characterized using molecular properties and fingerprints. TMap31 was used to characterize the data, a tree-based high-dimensional visualization tool, provides both local and distance structural cluster information. It clearly indicates that the data provided as a web server in this study are structurally diverse, with cell permeability assay information for the compounds also demonstrating significant diversity (**Fig. 3c**). The dataset consists of both non-peptide and semi-peptide macrocycles. The semi-peptides are situated on the right side of the TMap, exhibiting a higher fraction of sp3 carbons and relatively larger macrocyclic ring sizes compared to non-peptide macrocycles, which are predominantly represented by the macrocycles from the Rzepiela et al. dataset (*n = ~3500*).30

**Physicochemical property analysis:** To assess the diversity and molecular characteristics of non-peptidic macrocycles deposited in this work, we analyzed the distribution of key molecular features representing size (molecular weight [MW], topological polar surface area [TPSA]), hydrogen bonding characteristics (hydrogen bond acceptors [HBA], hydrogen bond donors [HBD]), lipophilicity (cLogP), flexibility (number of rotatable bonds [NRotB], Kier flexibility [φ]), and peptide nature (amide ratio). Most of the molecular descriptors adhere to the Lipinski26 and Veber27 rule drug-like cut-offs, particularly size, HBA, HBD, and NRotB, as the majority of this dataset was retrieved from the Rzepiela et al dataset30. 375 compounds, accounting for 8.9% of the macrocycles in the dataset, reside beyond the rule of five space (bRo5, MW > 500 Da and at least one of : MW 700-3000 Da, cLogP < 0 or > 7.5, HBD > 5, HBA > 10, TPSA > 200 Å2, NRotB > 20).3,5 These compounds exhibit the capability to modulate difficult-to-drug targets, including those with extensive, flat, or groove-shaped binding sites, as well as Protein-Protein Interactions (PPIs), while still allowing for oral administration.3

**Quantification of peptide and nonpeptide macrocycles.** Since no specific metric was used to quantify whether a macrocycle is of peptide or non-peptide nature in the literature before, we quantified the peptide nature of macrocycles based on the overall amide ratio (AR) (see Method section) and compared it among the macrocycles. The analysis (see **Fig. 4**) revealed that peptide macrocycles and non-peptide macrocycles could be well differentiated according to their AR. From the intersection point of the distribution curve, we computed MCC, sensitivity, specificity, and accuracy. Overall, the cut-off for non-peptide and peptide macrocycles was found to be ≤0.3 and ≥0.7, respectively. The overall quality of separation is remarkable, with an accuracy of 98%, sensitivity (peptide), specificity (non-peptide), and MCC found to be 0.98, 1.00, and 0.86, respectively. In addition, semi-peptide compounds were further assessed using these cut-offs, as the expected AR for semi-peptidic macrocycles was found to be ≥0.40, varying up to ≤0.85. In addition to these macrocyclic classes, some macrocycles could be categorized as semi-peptide macrocycles, typically characterized by an amide ratio between 0.4 and 0.7. This type of quantification would aid us in macrocycle design.

A collage of different types of molecules

Description automatically generated

**Figure 4**: a) Distribution of amide ratio of macrocycles (n = 7849) with representative examples are shown for low to high AR values. b) Principal component analysis (PCA) comparing the chemical space of peptides and non-peptides using the first two principal components (PCs), with descriptor contributions highlighted. Macrocycles are colored according to their amide ratio (AR) with blue to red circles. c) The chemical space of macrocycles reported in this study is depicted using the first two PCs, and colored according to their AR values.

### Usage Notes

### Here, we have developed a webserver for cell permeability data on non-peptidic macrocycles, comprising 5646 compounds collected from literature, patents, and public repositories to facilitate macrocyclic cell permeability modeling. The complete dataset can be accessed at <http://swemacrocycle.uu.se/> (Fig. 5a-c). The webserver offers three primary options for accessing macrocyclic cell permeability data: Browse, Download, and Analysis sections.

### Browse section: Users can access a comprehensive list of permeability datasets for macrocycles, featuring unique IDs, 2D structures, assay details, endpoints, values, units, standardized values (converted into logarithmic), and original sources. Clicking on each unique ID opens a separate window displaying the structure, permeability, and molecular features. Additionally, similar, or neighbouring molecules are listed based on the ‘macrocyclic core’ for comparison, enabling users to view similar compounds and their molecular characteristics. Users can also sort columns and filter rows to create customized subsets of macrocycles for analysis and download. The ‘Download’ section allows users to download the full dataset or subsets as CSV file with cell permeability information. Statistics section: Users can analyse both cell permeability data and molecular features. To enhance the user data analysis experience, we provide different analysis tools enabling users to select specific subsets and analyse their molecular features, followed by the option to download the specific subset for further analysis. In summary, the web server will be continuously updated with new data. Users can search for macrocycles by unique ID, name, assay type, molecular weight, endpoints, or a combination thereof for analysis within the web server and download the data as a CSV. The search results include detailed information about the structure, cell permeability, and molecular features, including peptide nature and original source.

A screenshot of a computer

Description automatically generated

**Figure 5.** Screenshots of “**Browse section**”, a) Data entry information for any macrocycles in the webserver, providing name, structural representation, permeability information, and molecular feature information. Additionally, for each entry, other permeability endpoints, if available, (b) are shown, and the 'Show Same-core Molecules' button (c) also lists similar core-containing molecules with molecular features.

### Code Availability

### All the data connected to this article is available without any restriction under http://swemacrocycle.com as webserver. All source code is also available on the GitHub (/macrocycle/cellperm/database/) with no restrictions to access.

### References

1 Wouters, O. J., McKee, M. & Luyten, J. Research and Development Costs of New Drugs-Reply. *JAMA* **324**, 518, doi:10.1001/jama.2020.8651 (2020).

2 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 Deliv Rev* **46**, 3-26, doi:10.1016/s0169-409x(00)00129-0 (2001).

3 Doak, B. C., Over, B., Giordanetto, F. & Kihlberg, J. Oral druggable space beyond the rule of 5: insights from drugs and clinical candidates. *Chem Biol* **21**, 1115-1142, doi:10.1016/j.chembiol.2014.08.013 (2014).

4 Poongavanam, V., Doak, B. C. & Kihlberg, J. Opportunities and guidelines for discovery of orally absorbed drugs in beyond rule of 5 space. *Curr Opin Chem Biol* **44**, 23-29, doi:10.1016/j.cbpa.2018.05.010 (2018).

5 Doak, B. C., Zheng, J., Dobritzsch, D. & Kihlberg, J. How Beyond Rule of 5 Drugs and Clinical Candidates Bind to Their Targets. *Journal of Medicinal Chemistry* **59**, 2312-2327, doi:10.1021/acs.jmedchem.5b01286 (2016).

6 Garcia Jimenez, D., Poongavanam, V. & Kihlberg, J. Macrocycles in Drug Discovery─Learning from the Past for the Future. *Journal of Medicinal Chemistry* **66**, 5377-5396, doi:10.1021/acs.jmedchem.3c00134 (2023).

7 Matsson, P., Doak, B. C., Over, B. & Kihlberg, J. Cell permeability beyond the rule of 5. *Adv Drug Deliv Rev* **101**, 42-61, doi:10.1016/j.addr.2016.03.013 (2016).

8 Over, B. *et al.* Structural and conformational determinants of macrocycle cell permeability. *Nat Chem Biol* **12**, 1065-1074, doi:10.1038/nchembio.2203 (2016).

9 Poongavanam, V. *et al.* Predicting the Permeability of Macrocycles from Conformational Sampling - Limitations of Molecular Flexibility. *J Pharm Sci* **110**, 301-313, doi:10.1016/j.xphs.2020.10.052 (2021).

10 Williams-Noonan, B. J. *et al.* Membrane Permeating Macrocycles: Design Guidelines from Machine Learning. *J Chem Inf Model* **62**, 4605-4619, doi:10.1021/acs.jcim.2c00809 (2022).

11 Bhardwaj, G. *et al.* Accurate de novo design of membrane-traversing macrocycles. *Cell* **185**, 3520-3532 e3526, doi:10.1016/j.cell.2022.07.019 (2022).

12 Merz, M. L. *et al.* De novo development of small cyclic peptides that are orally bioavailable. *Nature Chemical Biology*, doi:10.1038/s41589-023-01496-y (2023).

13 Hewitt, W. M. *et al.* Cell-permeable cyclic peptides from synthetic libraries inspired by natural products. *J Am Chem Soc* **137**, 715-721, doi:10.1021/ja508766b (2015).

14 Li, J. *et al.* CycPeptMPDB: A Comprehensive Database of Membrane Permeability of Cyclic Peptides. *J Chem Inf Model* **63**, 2240-2250, doi:10.1021/acs.jcim.2c01573 (2023).

15 Kim, S. *et al.* PubChem 2023 update. *Nucleic Acids Res* **51**, D1373-D1380, doi:10.1093/nar/gkac956 (2023).

16 Zdrazil, B. *et al.* The ChEMBL Database in 2023: a drug discovery platform spanning multiple bioactivity data types and time periods. *Nucleic Acids Res* **52**, D1180-D1192, doi:10.1093/nar/gkad1004 (2024).

17 RDKit (2010).

18 (Chemical Computing Group ULC, 910-1010 Sherbrooke St. W., Montreal, QC H3A 2R7, Canada, 2024).

19 Rossi Sebastiano, M. *et al.* Impact of Dynamically Exposed Polarity on Permeability and Solubility of Chameleonic Drugs Beyond the Rule of 5. *J Med Chem* **61**, 4189-4202, doi:10.1021/acs.jmedchem.8b00347 (2018).

20 Le Roux, A. *et al.* Structure-Permeability Relationship of Semipeptidic Macrocycles-Understanding and Optimizing Passive Permeability and Efflux Ratio. *J Med Chem* **63**, 6774-6783, doi:10.1021/acs.jmedchem.0c00013 (2020).

21 L'Exact, M. *et al.* Beyond Rule-of-five: Permeability Assessment of Semipeptidic Macrocycles. *Biochim Biophys Acta Biomembr* **1865**, 184196, doi:10.1016/j.bbamem.2023.184196 (2023).

22 Comeau, C. *et al.* Modulation of the Passive Permeability of Semipeptidic Macrocycles: N- and C-Methylations Fine-Tune Conformation and Properties. *J Med Chem* **64**, 5365-5383, doi:10.1021/acs.jmedchem.0c02036 (2021).

23 <https://www.rdkit.org>, R. O.-s. c.

24 Scalfani, V. F., Patel, V. D. & Fernandez, A. M. Visualizing chemical space networks with RDKit and NetworkX. *J Cheminform* **14**, 87, doi:10.1186/s13321-022-00664-x (2022).

25 Li, D. *et al.* ECharts: A declarative framework for rapid construction of web-based visualization. *Visual Informatics* **2**, 136-146, doi:<https://doi.org/10.1016/j.visinf.2018.04.011> (2018).

26 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. *Advanced Drug Delivery Reviews* **46**, 3-26, doi:10.1016/s0169-409x(00)00129-0 (2001).

27 Veber, D. F. *et al.* Molecular Properties That Influence the Oral Bioavailability of Drug Candidates. *Journal of Medicinal Chemistry* **45**, 2615-2623, doi:10.1021/jm020017n (2002).

28 Poongavanam, V. *et al.* Conformational Sampling of Macrocyclic Drugs in Different Environments: Can We Find the Relevant Conformations? *ACS Omega* **3**, 11742-11757, doi:10.1021/acsomega.8b01379 (2018).

29 Sethio, D. *et al.* Simulation Reveals the Chameleonic Behavior of Macrocycles. *J Chem Inf Model* **63**, 138-146, doi:10.1021/acs.jcim.2c01093 (2023).

30 Rzepiela, A. A., Viarengo-Baker, L. A., Tatarskii, V., Kombarov, R. & Whitty, A. Conformational Effects on the Passive Membrane Permeability of Synthetic Macrocycles. *J Med Chem* **65**, 10300-10317, doi:10.1021/acs.jmedchem.1c02090 (2022).

31 Probst, D. & Reymond, J.-L. Visualization of very large high-dimensional data sets as minimum spanning trees. *Journal of Cheminformatics* **12**, 12, doi:10.1186/s13321-020-0416-x (2020).

### 

### Acknowledgements

The work was funded by a grant from the Swedish Research Council (Grant No. 2021-04747). The authors are grateful to Saw Simeon for assisting with TMap figure generation and thank Mikael Österberg, Uppsala University for IT support. We also thank OpenEye scientific software and ChemAxon for providing academic licenses.

### Author contributions

FQ: Methodology, validation, formal analysis, data curation; DDC: methodology, validation, formal analysis; RL: funding acquisition, writing – review; JK: conceptualization, formal analysis, data curation, writing – original draft & review, project supervision, funding acquisition; VP: conceptualization, methodology, formal analysis, data curation, visualization, writing – original draft & review, and project supervision.

### Competing interests

The authors declare that they have no known competing financial interests.