IBMC

XXXI Symposium on Bioinformatics and Computer-Aided Drug Discovery (BCADD-2025)

Emerging Challenges and Opportunities for In Silico Drug Discovery

Way2Drug PREDICTIVE SERVICES
Understanding Chemical-Biological Interactions

TOOL FOR DIVERSITY VISUALIZATION ON THE LEVEL OF MOLECULAR SCAFFOLDS, TDV, CHEMICAL DATA AT GLANCE

Pavel V. Pogodin PhD, scientist at the LSFBD, IBMC

2025-10-22, Moscow

Scaffolds

Scaffold is the core (rings + linkers between them) of the chemical structure,

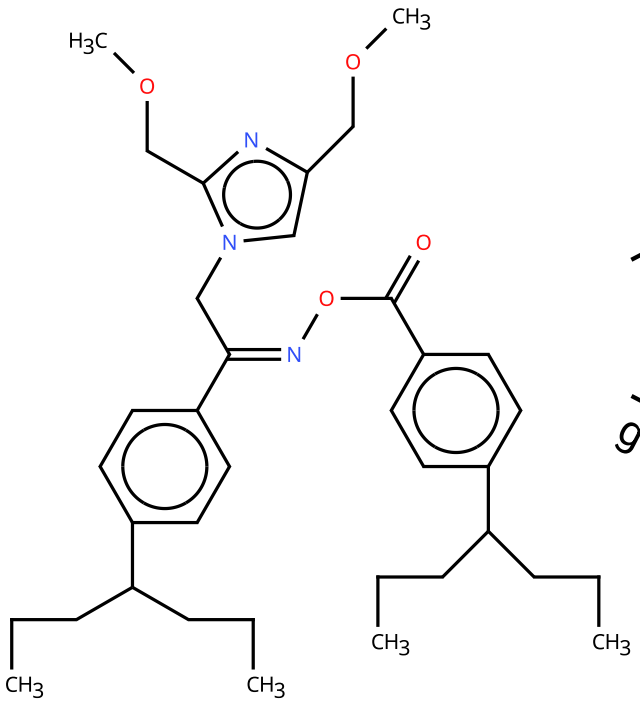
- which, to a large extent, defines **molecule's** geometry,
- which greatly affects the possibility for the **molecule** to interact with the components of biological system and thus,
- to some extent defines the activity spectra[1] of the **molecule** in the human body, for example.

Scaffolds

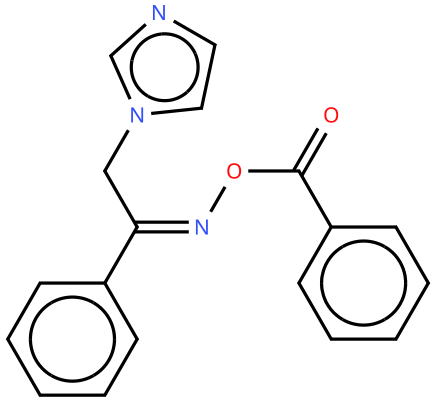
This concept was introduced to the broad scientific community as Molecular Framework by medicinal & computational chemists[2], thus **molecule** here is rather a **drug-like molecule**.

At the moment it is possible to calculate scaffolds using various cheminformatics tools, RDKit for example[3], and meaning of this term could slightly vary from tool to tool.

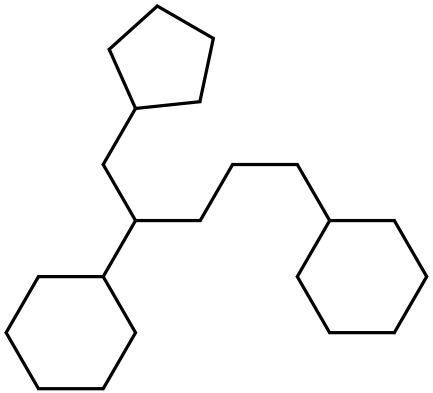
RDKit scaffolds



scaffold

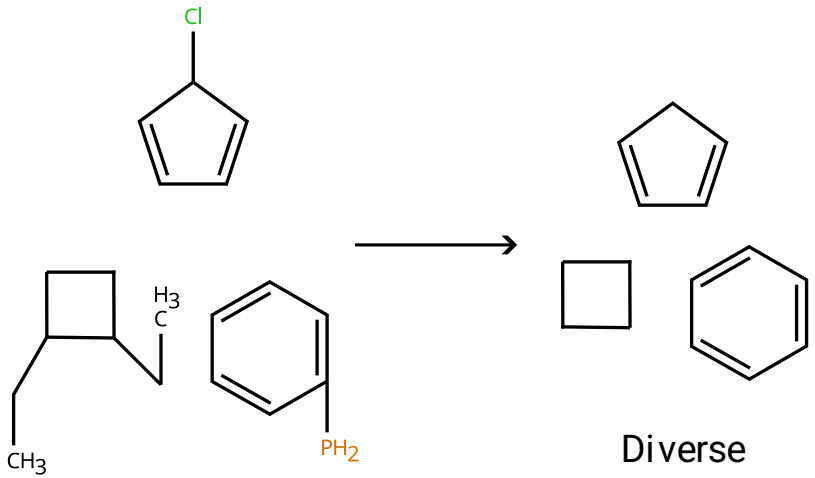


generic scaffold



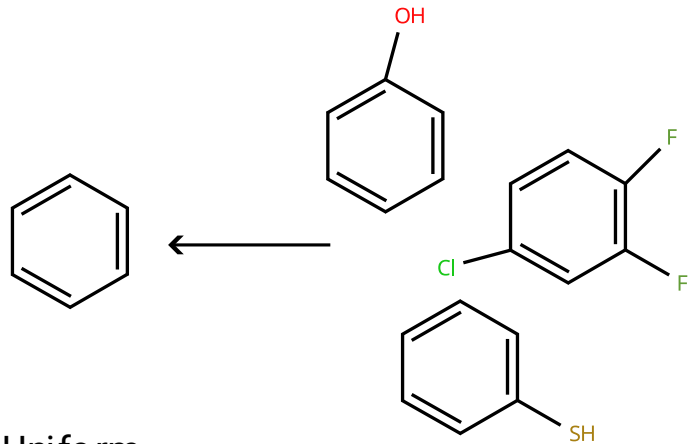
Scaffolds are molecules too, they should be valid

Diversity



Diverse

Uniform



In the context of this study, diversity is the compositional complexity of the set of chemical structures assessed as the number and weight of the scaffolds describing them.

Diversity

- One can see the analogy with the concept of biodiversity assessed through the number and weight of the species in the ecosystem and dive deeper[4–6].
- Others can see the connections with the molecular complexity extended from the level of distinct structures to the level of the set of structures[7].

Diversity

In short, this topic is hard, but useful:

- there are several ways to think of and assess the diversity
- all of them are somehow different from each other and that could have some practical consequences
- In the meantime, knowing the content of the set of chemical structures is useful, it can help to select appropriate modeling strategy and rationally select the specific set to work with among the several available.

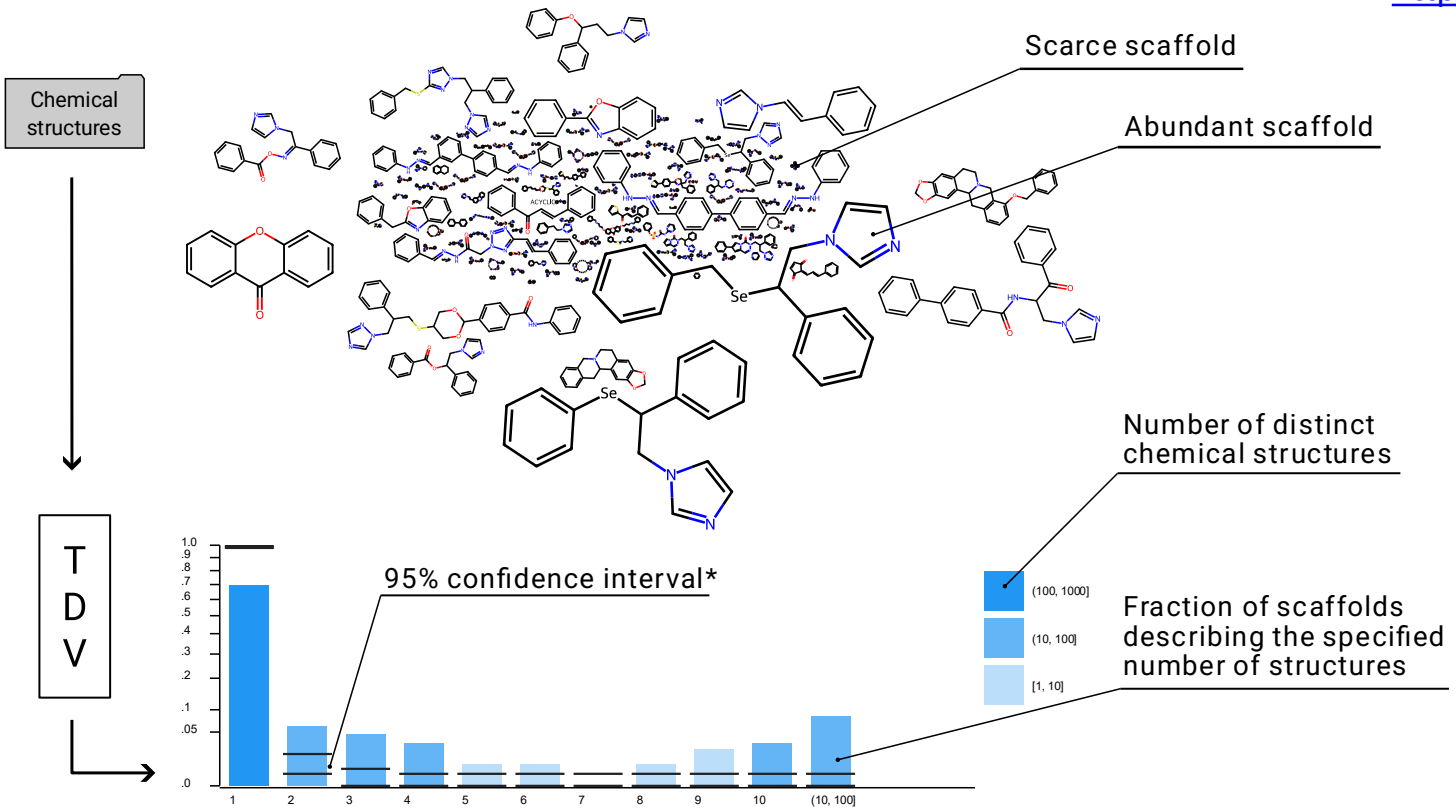
Diversity

Thus, diversity assessment should benefit from the actionable visualization[8–13].

Tool for chemical diversity visualization on the level of molecular scaffolds

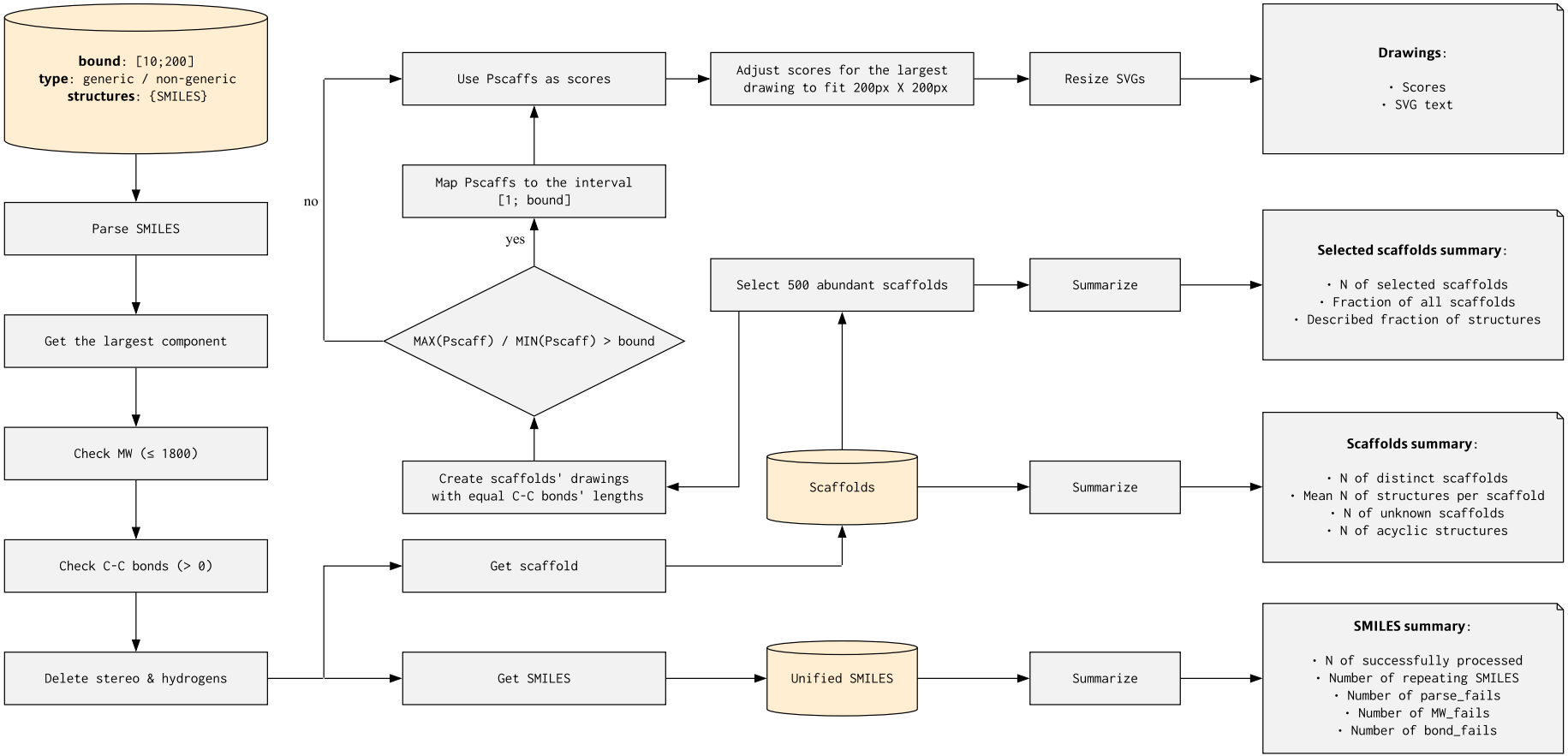
The dataset of C. albicans inhibitors
Constructed using ChEMBL data

https://way2drug.com/tdv/



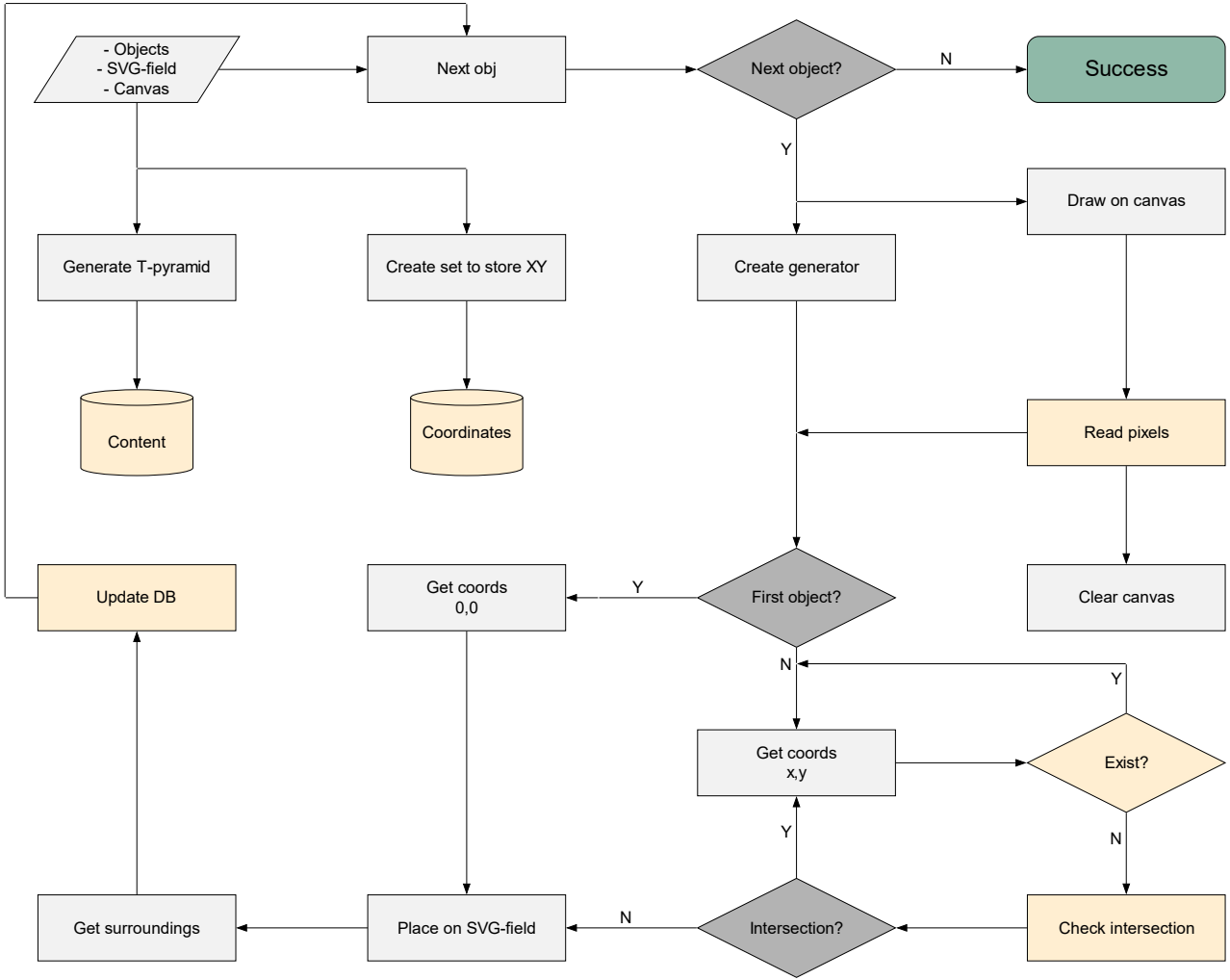
*Confidence intervals were calculated via bootstrap for the random sample of PubChem and ChEMBL data combined

How it works



on the server

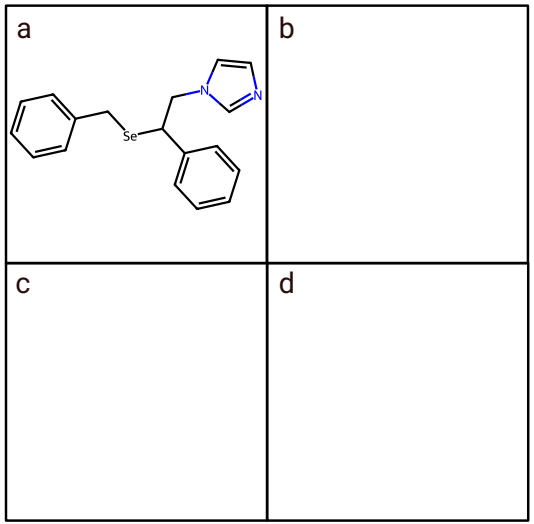
How it works



in the browser

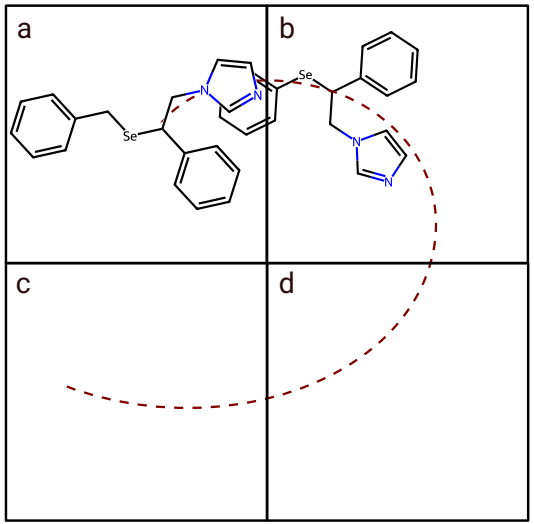
How it works

1. Place the first, largest, scaffold



OK

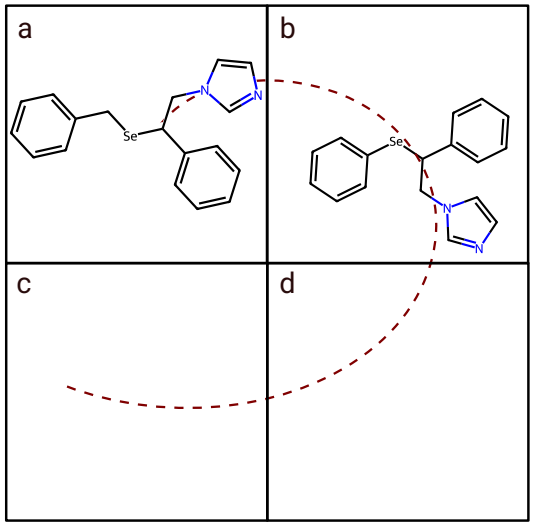
2. Place the second scaffold with an offset



- identify quadrants occupied by the scaffold (a and b)
- check XY coordinates for intersection

not OK

3. Move the second scaffold



- identify quadrants occupied by the scaffold (b)
- check XY coordinates for intersection

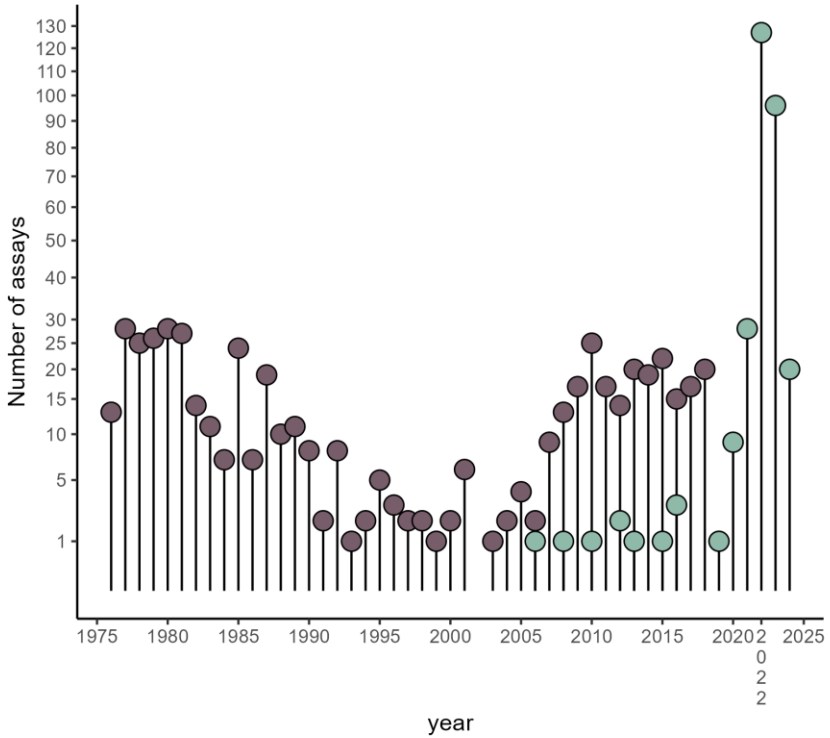
OK

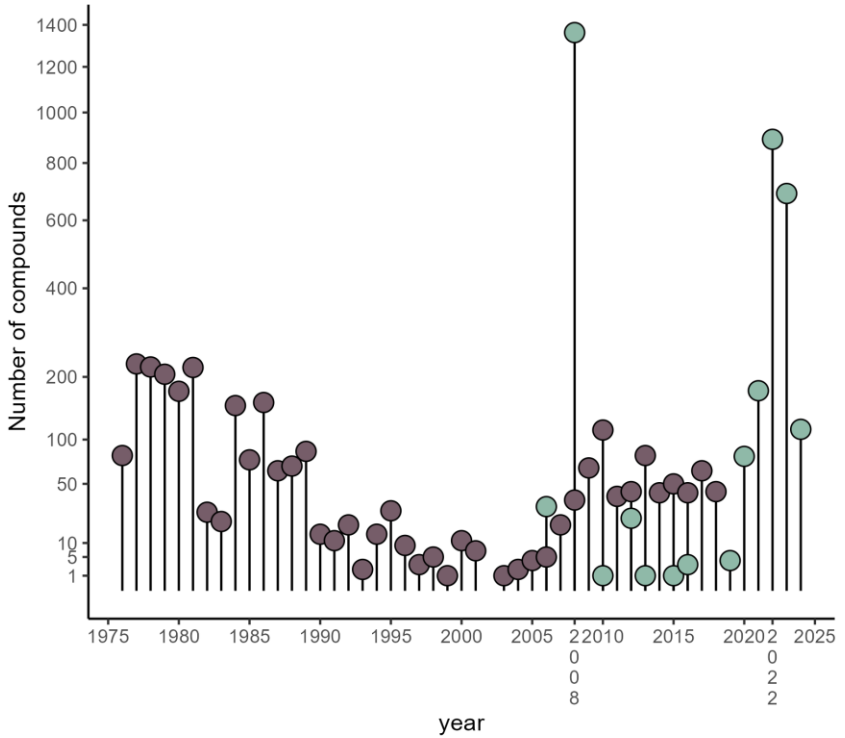
Proceed with the next scaffold

Case study of the ChEMBL data on toxicity[14], RSF № 25-15-00300

- Obtain ChEMBL[15] data related to the pre-clinical assessment of toxicity: Chemical Structures of compound having LD_{50} (mouse) and / or CC_{50} (cells)
- Visualize the diversity on the level of scaffolds using TDV and on the level of MNA descriptors[16] using extended UMAP[17–19]

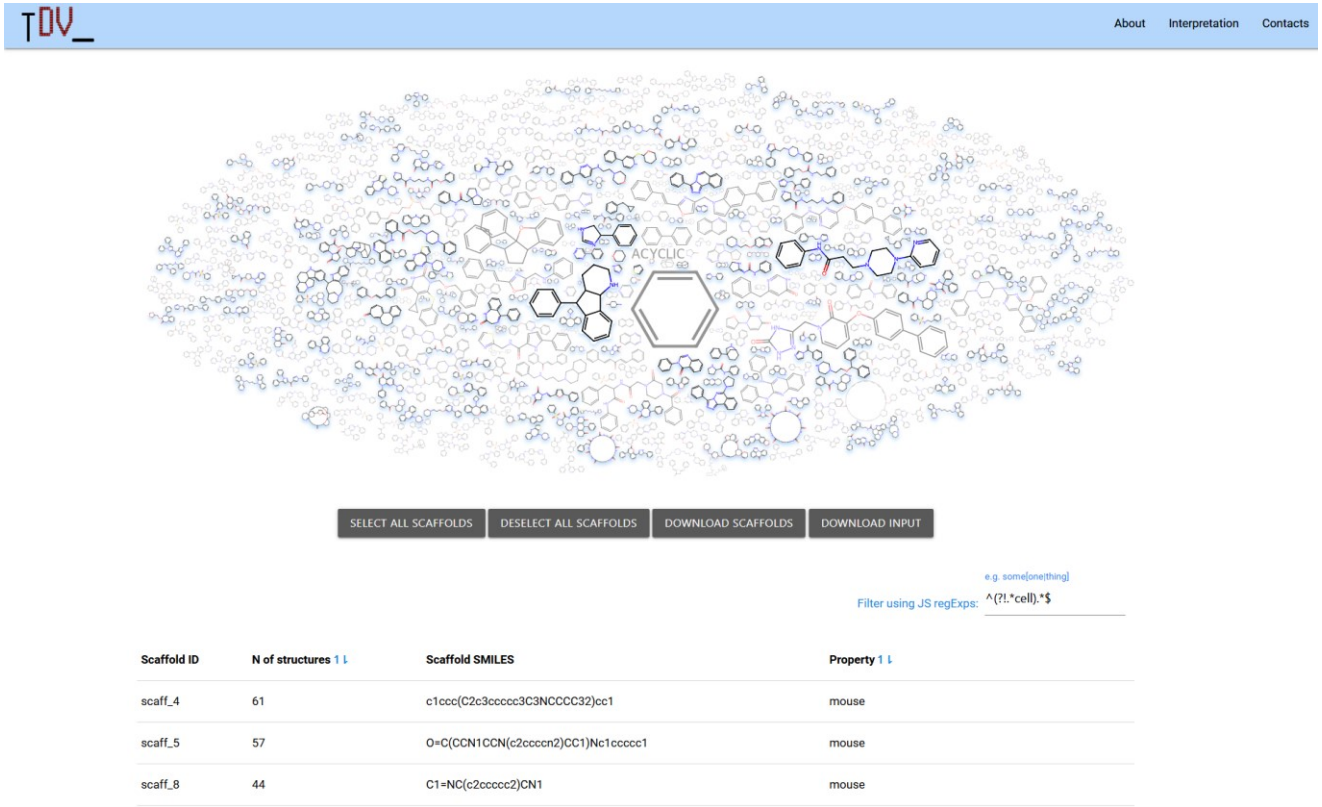
Data





Mouse probably leaves the freely available LD_{50} studies for good

Elements of the scaffold’s diversity analysis

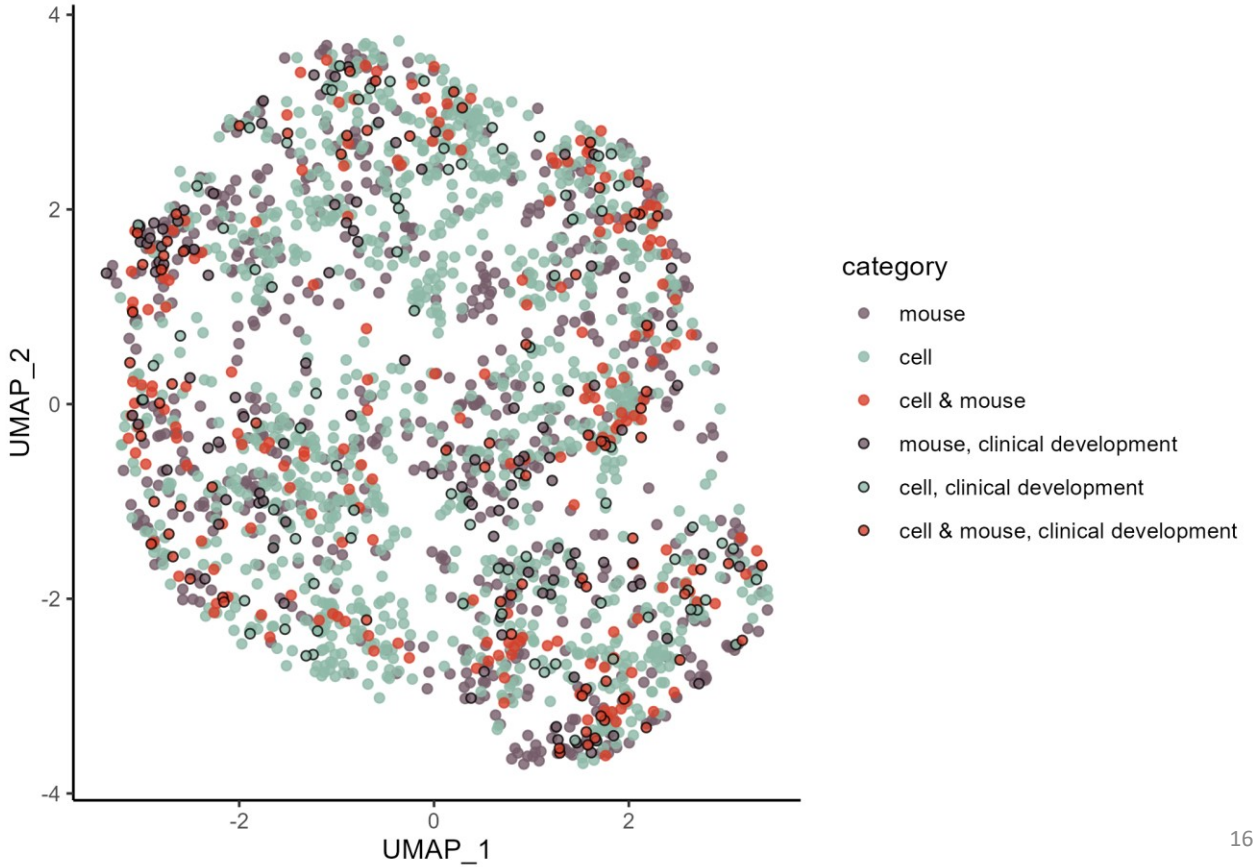


Scaffolds describing compounds studied using only mouse and not cells are highlighted

Diversity analysis on the level of MNA descriptors

- Describe compounds using MNA
- Select appropriate number of descriptors allowing to distinguish between the compounds
- Extend the set appropriately
- UMAP
- Plot the results

Chemical structures tested using cells cover all the available space of descriptors



Materials

TDV:

- https://github.com/RSF-23-73-01058/
- https://way2drug.com/tdv/

LD_{50} and CC_{50} study:

- https://github.com/RSF-25-15-00300-LD50-cytotoxicity/ChEMBL35_compare_mouseLD50-cellsCC50

Key takeaways

- Working with chemical data is important for the clarity of the various studies involving them

- Diversity of the sets of chemical compounds could be visualized using TDV freely available in the Internet via Way2Drug platform: https://way2drug.com/tdv/

Acknowledgements

- TDV development was supported by the Russian Science Foundation, grant № 23-73-01058
- LD_{50} and CC_{50} analysis was conducted in the framework of the project supported by the Russian Science Foundation, grant № 25-15-00300
- To the G.S. Malakhov (FBB, MSU; LSFBDD, IBMC), who contributed greatly to the TDV development and to the discussions of related topics
- To the Y.V. Polovets, who prepared the C. albicans dataset and observed the deviations in the global data structure after UMAP in the course of her graduation project (RNRMU)
- To the D.S. Druzhilovsky - for the TDV placement on the Way2Drug platform
- To the head of LSFBDD, full member of RAS, prof. V.V. Poroikov and all the colleagues from LSFBDD and IBMC and scientific community in general - for the fruitful discussions

THANK YOU FOR YOUR ATTENTION

References

(1) Lagunin, A.; Stepanchikova, A.; Filimonov, D.; Poroikov, V. PASS: Prediction of Activity Spectra for Biologically Active Substances. *Bioinformatics* **2000**, *16* (8), 747–748.

(2) Bemis, G. W.; Murcko, M. A. The Properties of Known Drugs. 1. Molecular Frameworks. *Journal of medicinal chemistry* **1996**, *39* (15), 2887–2893.

(3) Landrum, G. Rdkit Documentation. *Release* **2013**, *1* (1-79), 4.

(4) Patil, G. P.; Taillie, C. Diversity as a Concept and Its Measurement. *Journal of the American statistical Association* **1982**, *77* (379), 548–561.

(5) Roswell, M.; Dushoff, J.; Winfree, R. A Conceptual Guide to Measuring Species Diversity. *Oikos* **2021**, *130* (3), 321–338.

(6) Chao, A.; Chiu, C.-H.; Villéger, S.; Sun, I.-F.; Thorn, S.; Lin, Y.-C.; Chiang, J.-M.; Sherwin, W. B. An Attribute-Diversity Approach to Functional Diversity, Functional Beta Diversity, and Related (Dis) Similarity Measures. *Ecological monographs* **2019**, *89* (2), e01343.

(7) Bonchev, D.; Rouvray, D. *Complexity: Introduction and Fundamentals*; CRC Press, 2003.

(8) Schneider, T. D.; Stephens, R. M. Sequence Logos: A New Way to Display Consensus Sequences. *Nucleic acids research* **1990**, *18* (20), 6097–6100.

(9) Viegas, F. B.; Wattenberg, M.; Feinberg, J. Participatory Visualization with Wordle. *IEEE transactions on visualization and computer graphics* **2009**, *15* (6), 1137–1144.

(10) Ertl, P.; Rohde, B. The Molecule Cloud-Compact Visualization of Large Collections of Molecules. *Journal of Cheminformatics* **2012**, *4* (1), 12.

(11) González-Medina, M.; Prieto-Martínez, F. D.; Owen, J. R.; Medina-Franco, J. L. Consensus Diversity Plots: A Global Diversity Analysis of Chemical Libraries. *Journal of Cheminformatics* **2016**, *8* (1), 63.

(12) Gaspar, H. A.; Baskin, I. I.; Varnek, A. Visualization of a Multidimensional Descriptor Space. In *Frontiers in molecular design and chemical information science-herman skolnik award symposium 2015: Jürgen bajorath*; ACS Publications, 2016; pp 243–267.

(13) Probst, D.; Reymond, J.-L. Visualization of Very Large High-Dimensional Data Sets as Minimum Spanning Trees. *Journal of Cheminformatics* **2020**, *12* (1), 12.

(14) scientists, various young. XI Молодежная Конференция ИОХ РАН.

(15) Zdrazil, B.; Felix, E.; Hunter, F.; Manners, E. J.; Blackshaw, J.; Corbett, S.; De Veij, M.; Ioannidis, H.; Lopez, D. M.; Mosquera, J. F.; others. The ChEMBL Database in 2023: A Drug Discovery Platform Spanning Multiple Bioactivity Data Types and Time Periods. *Nucleic acids research* **2024**, *52* (D1), D1180–D1192.

(16) Filimonov, D.; Poroikov, V.; Borodina, Y.; Gloriozova, T. Chemical Similarity Assessment Through Multilevel Neighborhoods of Atoms: Definition and Comparison with the Other Descriptors. *Journal of chemical information and computer sciences* **1999**, *39* (4), 666–670.

(17) McInnes, L.; Healy, J.; Melville, J. Umap: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv preprint arXiv:1802.03426* **2018**.

(18) Konopka, T. Umap: Uniform Manifold Approximation and Projection.

(19) Pogodin, P. Local and Global Data Structure Preservation in Dimensionality Reduction: A Case Study of the UMAP Algorithm. In *New horizons of applied mathematics*; 2025; Vol. 2, pp 105–107.

(20) Team, R. C. R Language Definition. *Vienna, Austria: R foundation for statistical computing* **2000**, *3* (1), 116.

(21) Wickham, H.; Averick, M.; Bryan, J.; Chang, W.; McGowan, L. D.; François, R.; Grolemund, G.; Hayes, A.; Henry, L.; Hester, J.; others. Welcome to the Tidyverse. *Journal of open source software* **2019**, *4* (43), 1686.

(22) Guha, R. Chemical Informatics Functionality in r. *Journal of Statistical Software* **2007**, *18*, 1–16.

(23) Kursa, M. B. Praznik: High Performance Information-Based Feature Selection. *SoftwareX* **2021**, *16*, 100819.