Supplementary Information
for the article

Digitization of molecular complexity with machine learning

Andrei S. Tyrin,[#] Daniil A. Boiko,[#] Nikita I. Kolomoets, Valentine P. Ananikov*

Zelinsky Institute of Organic Chemistry, Russian Academy of Sciences, Leninsky prospekt 47, Moscow, 119991, Russia; http://AnanikovLab.ru
*e-mail: val@ioc.ac.ru; [#]equal contributions

Contents

1. Dataset

As was mentioned, the collected dataset consisted of three phases (consult the paper for detailed description and motivation for each of them). Using the quality controls and group membership as filtering criteria, we obtained the following count distribution for each of the phases:

1) **Absolute values phase** — 10978 labels used.
2) **Active learning phase** — 94810 labels used.
3) **Ranking phase** — 90295 labels used.

The sizes of three phases compared in the Figure S1.

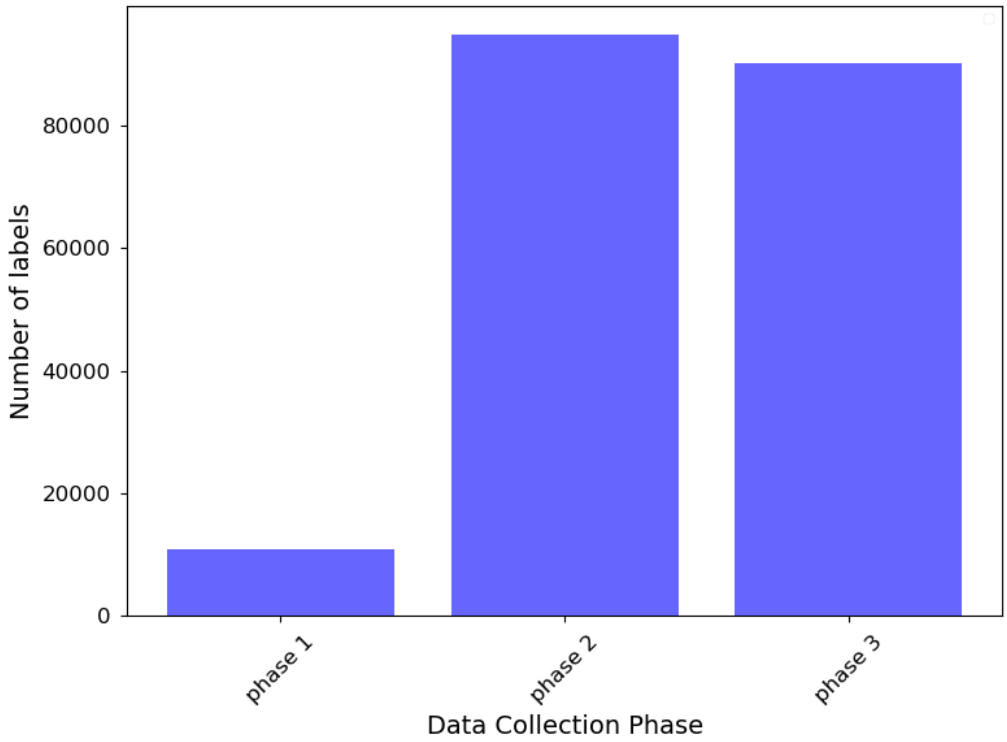


Figure S1. Distribution of votes number across different phases.

Active learning allowed to collect labels for 164017 molecules, which is a significant increase compared to other data-based approaches tackling the molecular complexity determination.

Each expert was asked their education level. The results are presented in Figure S2. As can be seen, the largest group of experts consisted of PhD students pursuing organic chemistry as their domain of expertise followed by experts who already obtained PhD in organic chemistry.

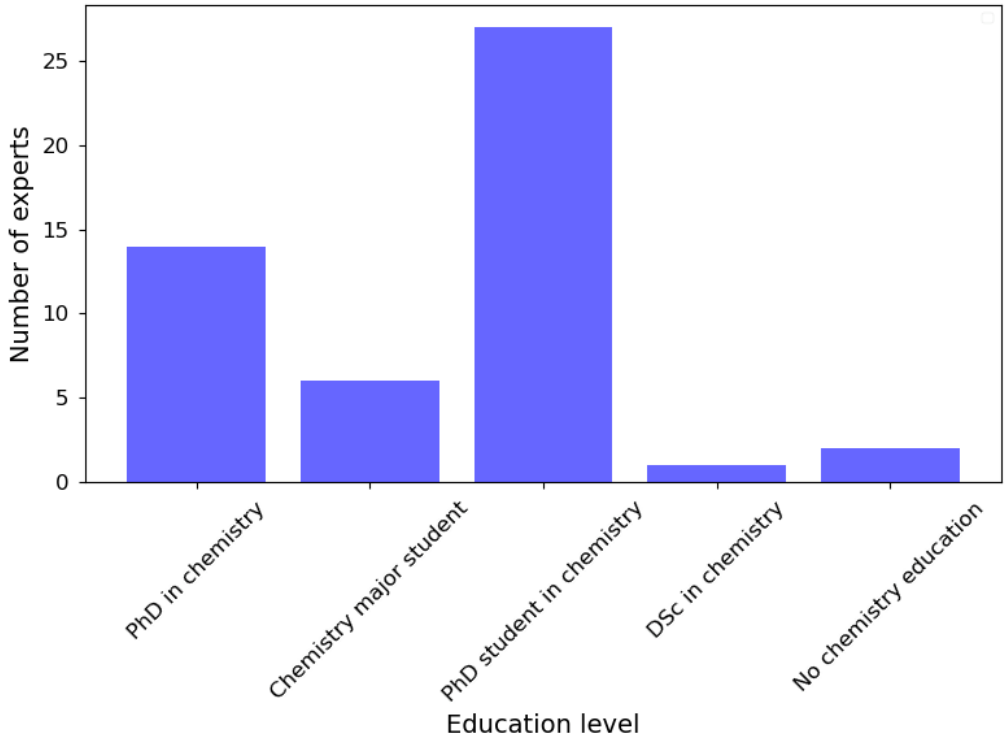


Figure S2. Education level among the experts whose labels served as labels for the ranking model training.

2. Active Learning

After collecting the initial set of labels with randomly sampled groups of molecules the active learning was employed [links]. The reason for integrating active learning into the data collection pipeline is to cover as much of the chemical space by using as little of human effort as possible. During the most productive days of data collection, the ensemble of ranking models was retrained nightly using the already collected data, after which it was used to select a batch of unlabeled molecules using uncertainty in the ranking score as a criterion. After this step, selected molecules were grouped into 5 clusters (corresponding to 5 complexity thresholds), and within each cluster most similar

molecules were grouped into samples that were provided to the experts on the following day. This step, thus achieved two goals: covering the chemical space as efficient as possible as well as collecting the labels for the molecules that the model considers highly similar.

3. Model Selection

Although having different data collection phases is beneficial simple merging of them for the subsequent training might be hard due to the different sizes and types of molecules in each phase. For instance, the second phase is based on the data collected during the first phase and data collected during active learning. Therefore, we used the group weight parameter between different phases of the GBDTs model as a hyperparameter for the selection of best model. Other parameters included:

```
"loss_function": "YetiRank",
"iterations": 1000,
"depth": 8,
"learning_rate": 0.05,
"bootstrap_type": "MVS",
```

We found that changing them didn't significantly change the performance of the model on the test dataset that was obtained by random split. Therefore, the main hyperparameter that was used corresponded to the group weight. The weight was discretized up to 0.1 thus resulting in 66 different models. Pair Accuracy and Function Group Test (FGT) was used as selection criteria. The functional groups included: phenyl, methyl, amine, hydroxyl, carboxyl, aldehyde, t-Bu, isopropyl, nitril, sulfo, fluorine, chlorine, bromine, iodine, and methoxy groups.

By equipping this approach for the hyperparameter tuning we were able to combine the data collected in different phases by maximizing the ranking quality as well as having the meaningful MC predictions validated by FGT. The detailed results of the conducted analysis are presented in the Figure S3. As was mentioned in the main part of the article, the selected phase weights correspond to 70%, 10%, and 20% distributed

between the first, second, and third phases respectively. This distribution corresponds to the optimal joint performance from both PA and FGT perspectives.

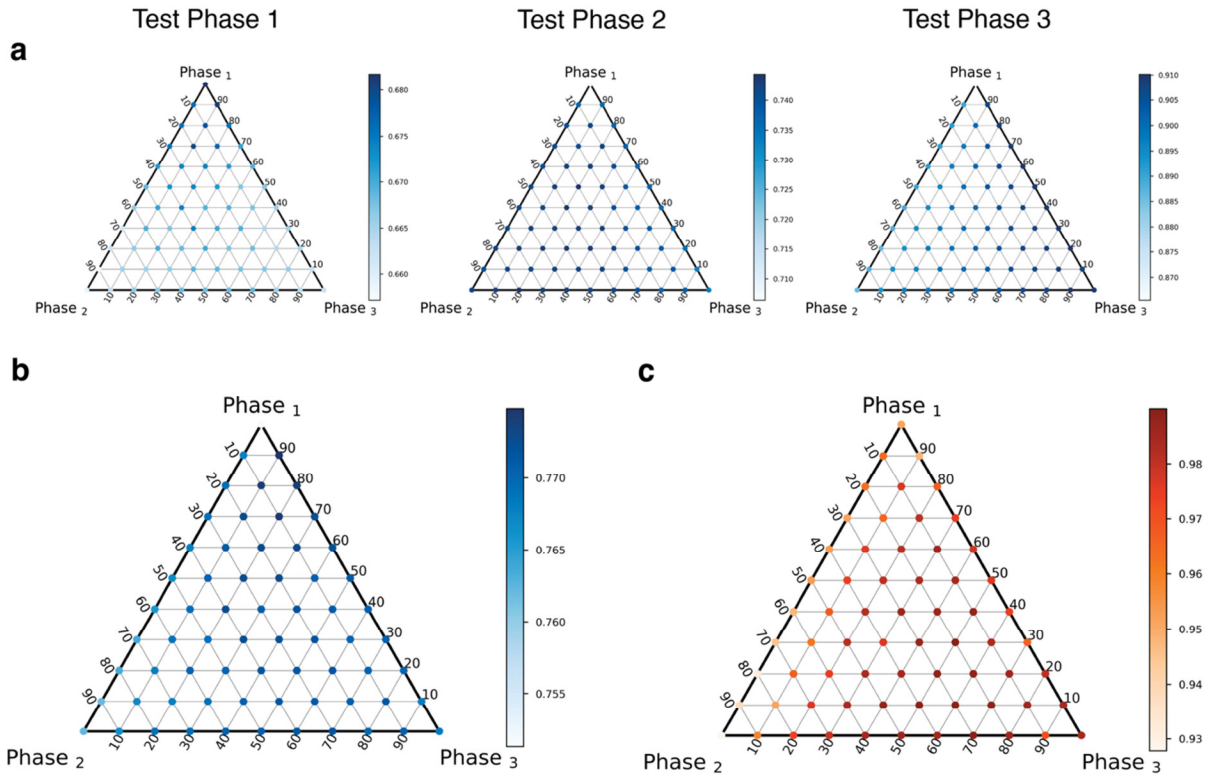


Figure S3. Ternary diagram visualization for the performance metrics. (a) pair accuracy on different test phases obtained by random train/test split; (b) averaged pair accuracy for different phases; (c) FGT results (the higher score/darker color shade is better).

4. Interpreting molecular complexity

In addition to Fig. 1a, the complete SHAP beeswarm plot is given in Fig. S3.

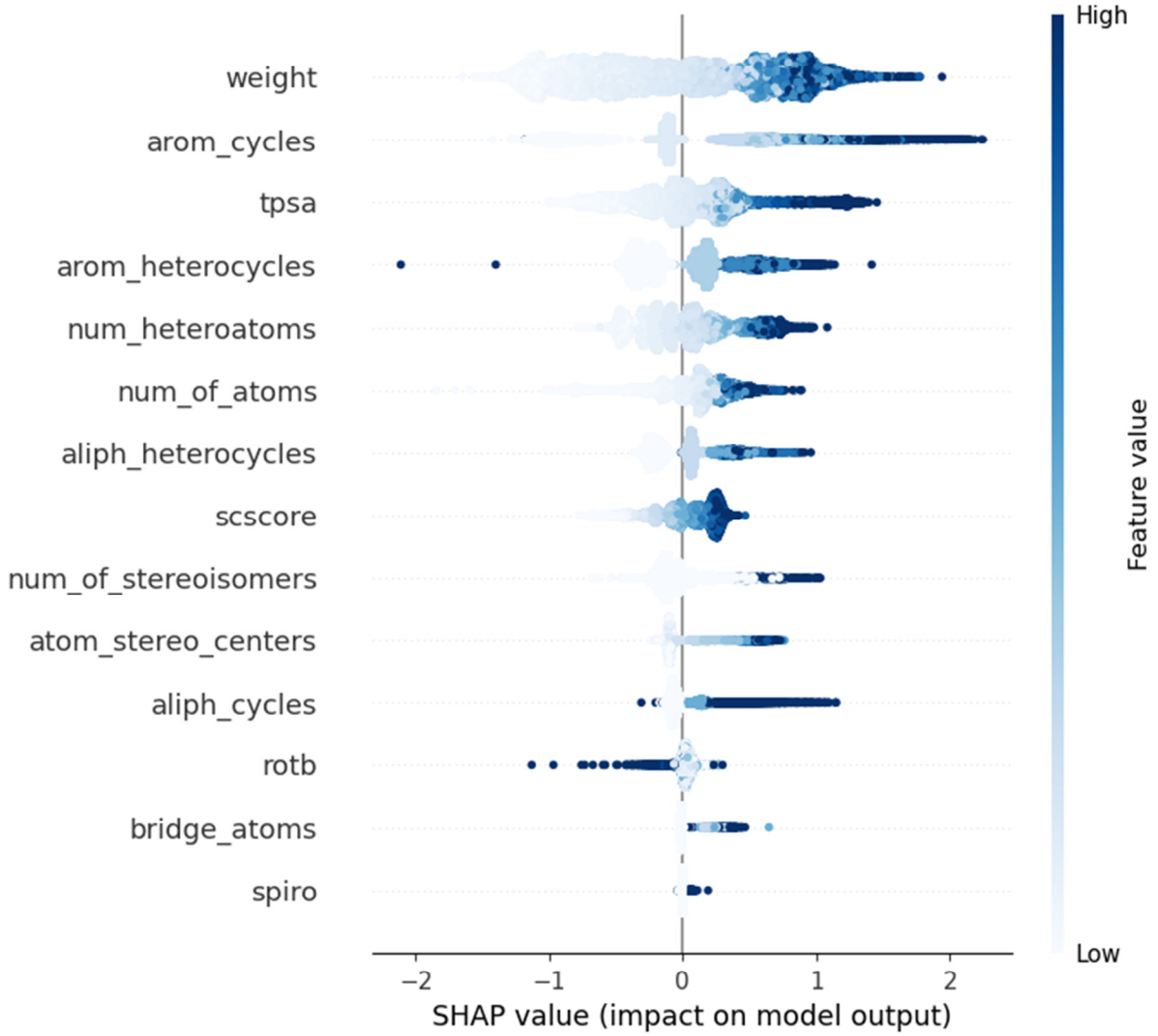


Figure S4. SHAP beeswarm plot presenting the feature importance of all the features.

To further interpret the relationship between the molecular complexity and other molecular features that were used for the training of the machine learning model, Fig. S5 was made.

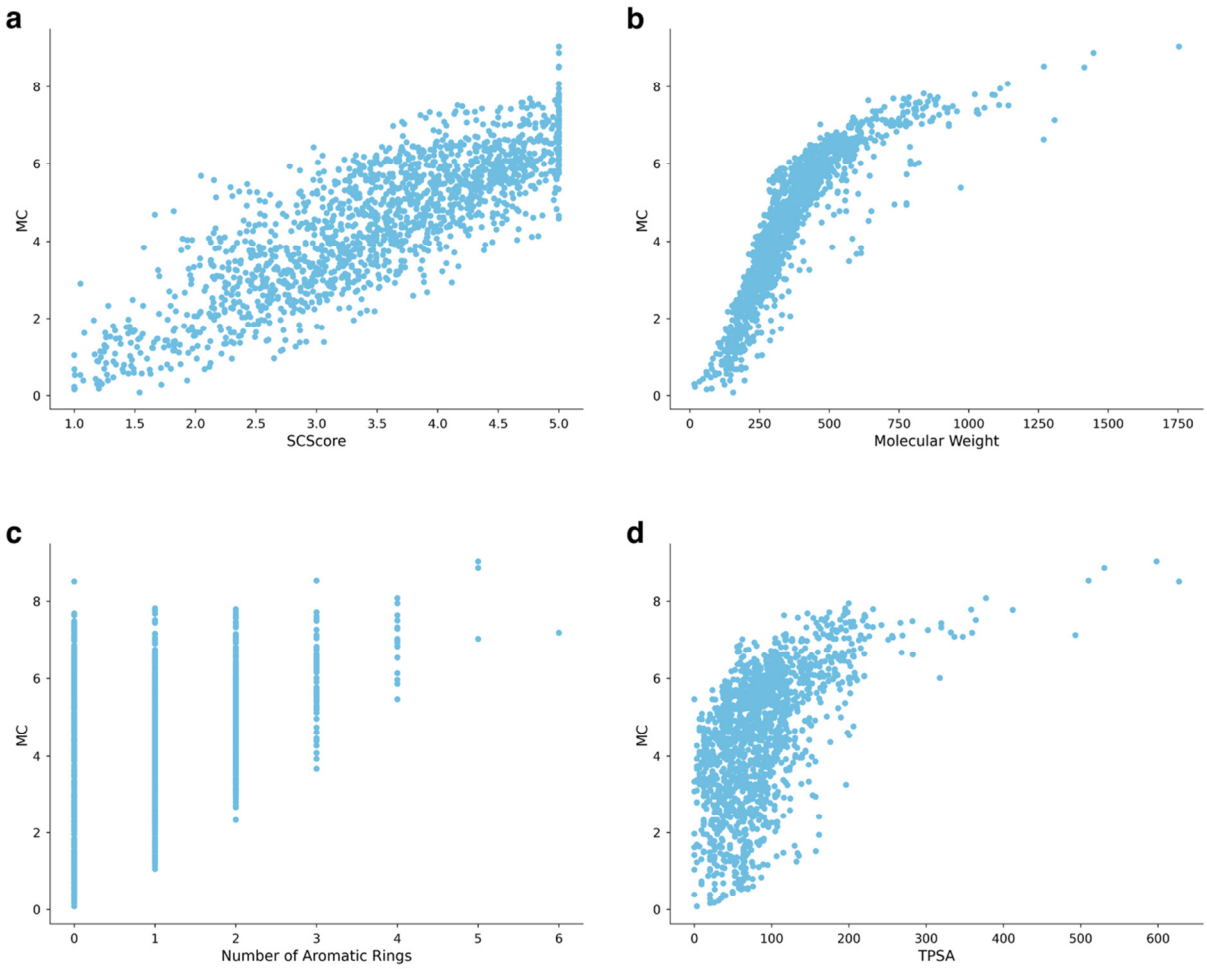


Figure S5. MC plotted against the (a) SCScore, (b) molecular weight, (c) number of aromatic rings, and (d) topological polar surface area. Molecules were taken from the ChEMBL database.

Fig. S5 allows to explore further what guides scientists during the analysis of molecules from the MC perspective.

5. FDA drugs results

Additionally to the 2D histogram presented in the Fig. 3a of the main text, we provide a 3D histogram (Fig. S6) to provide an additional perspective on the molecular complexity evolution of small molecule drugs. Additionally, Fig. S6 contains the results of the Mann-Kendall test, indicating clear growth of the molecular complexity over years.

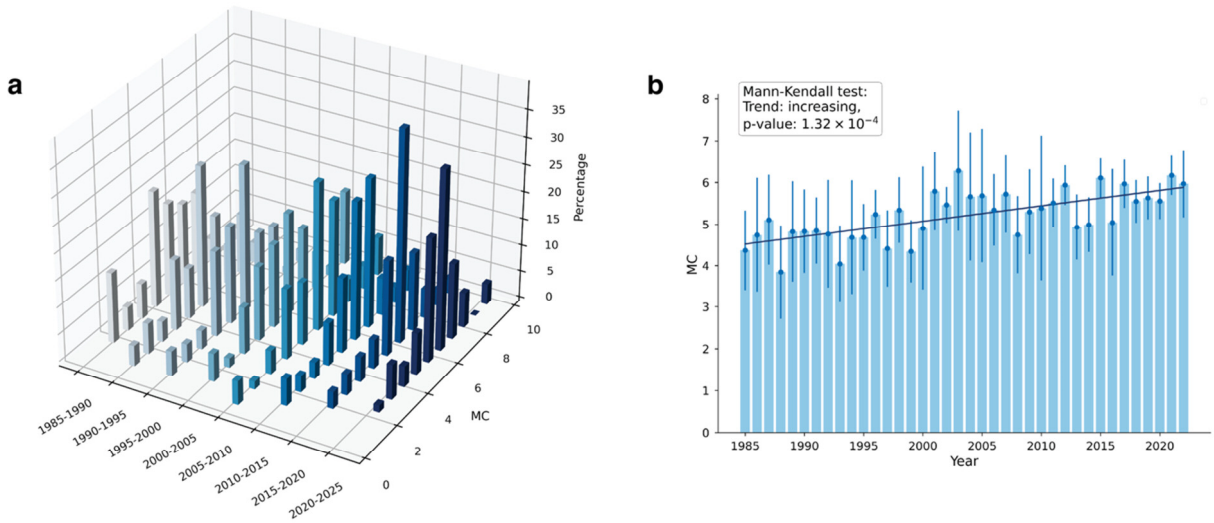


Figure S6. (a) 3D histogram illustrating the MC distribution of drugs over time; (b) histogram with the results of Mann-Kendall test;

In addition to the molecules of the median molecular complexity per 5 years intervals illustrated in the main text, we provide the MC-annotated molecular structures in Fig. S7.

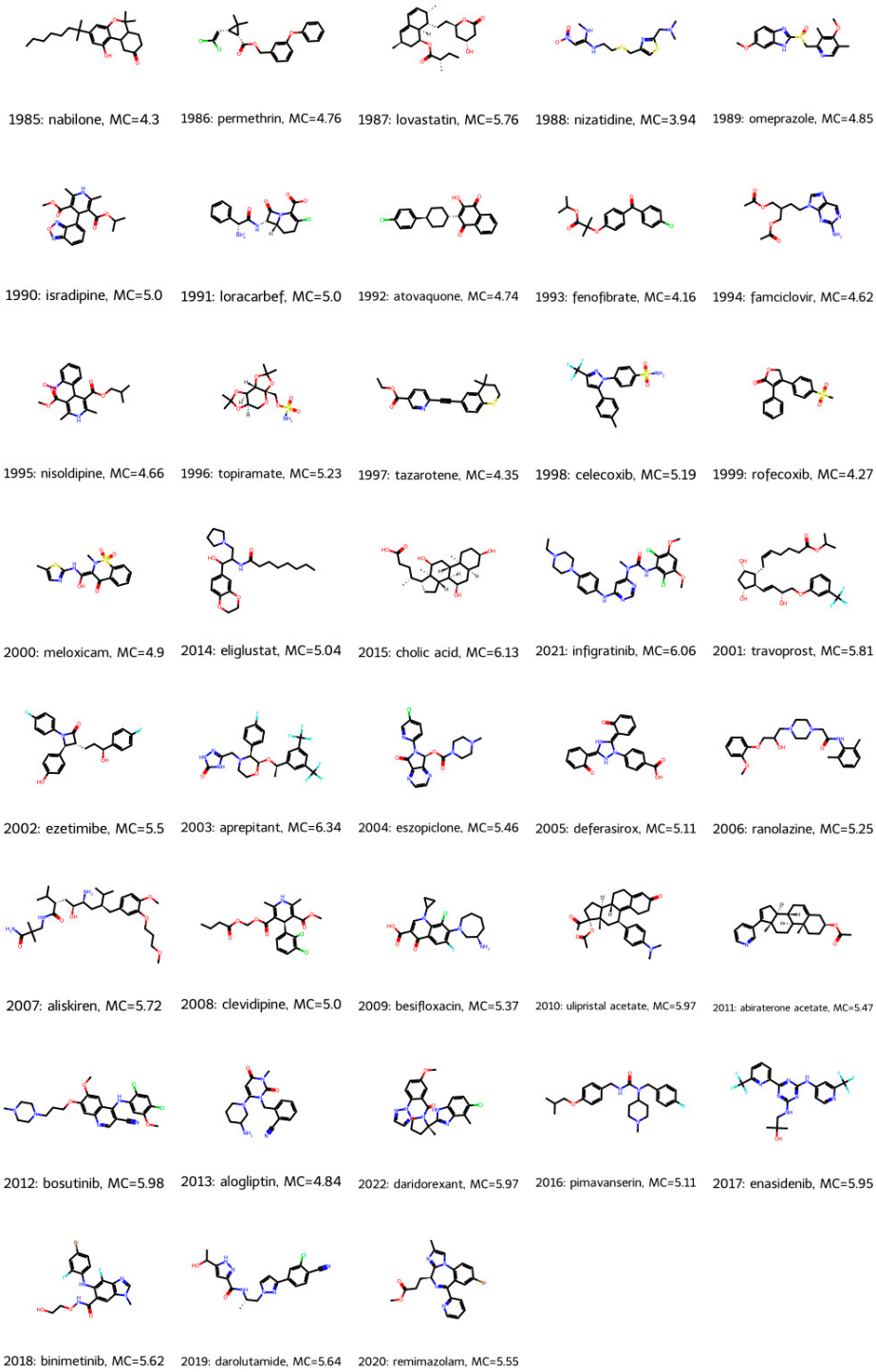


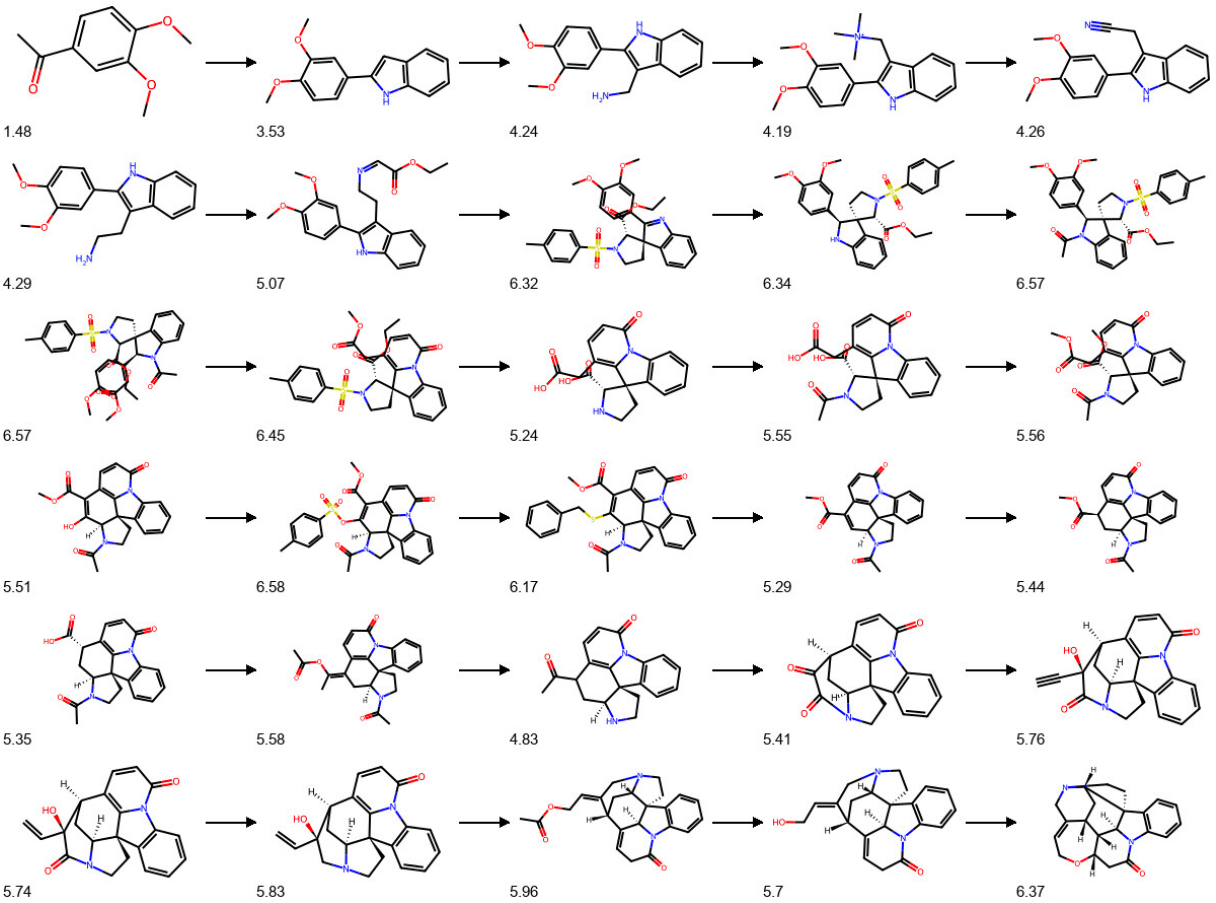
Figure S7. FDA approved drugs with closest to the average molecular complexity value per each year.

6. Total syntheses with molecular complexity values

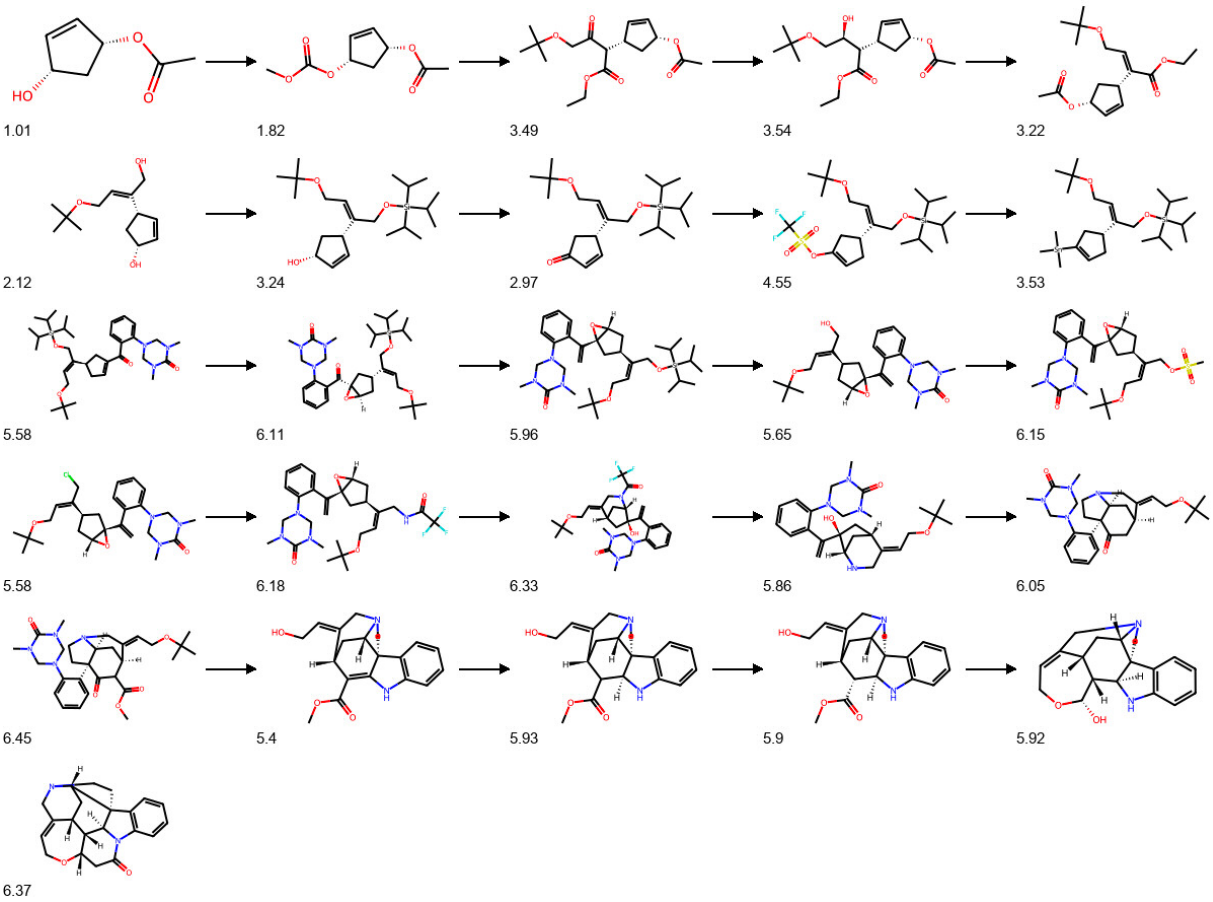
The schemes below contain the synthetic route strychnine and artemisinin. All the schemes contain the intermediates with the biggest molecular complexity values.

1. Strychnine

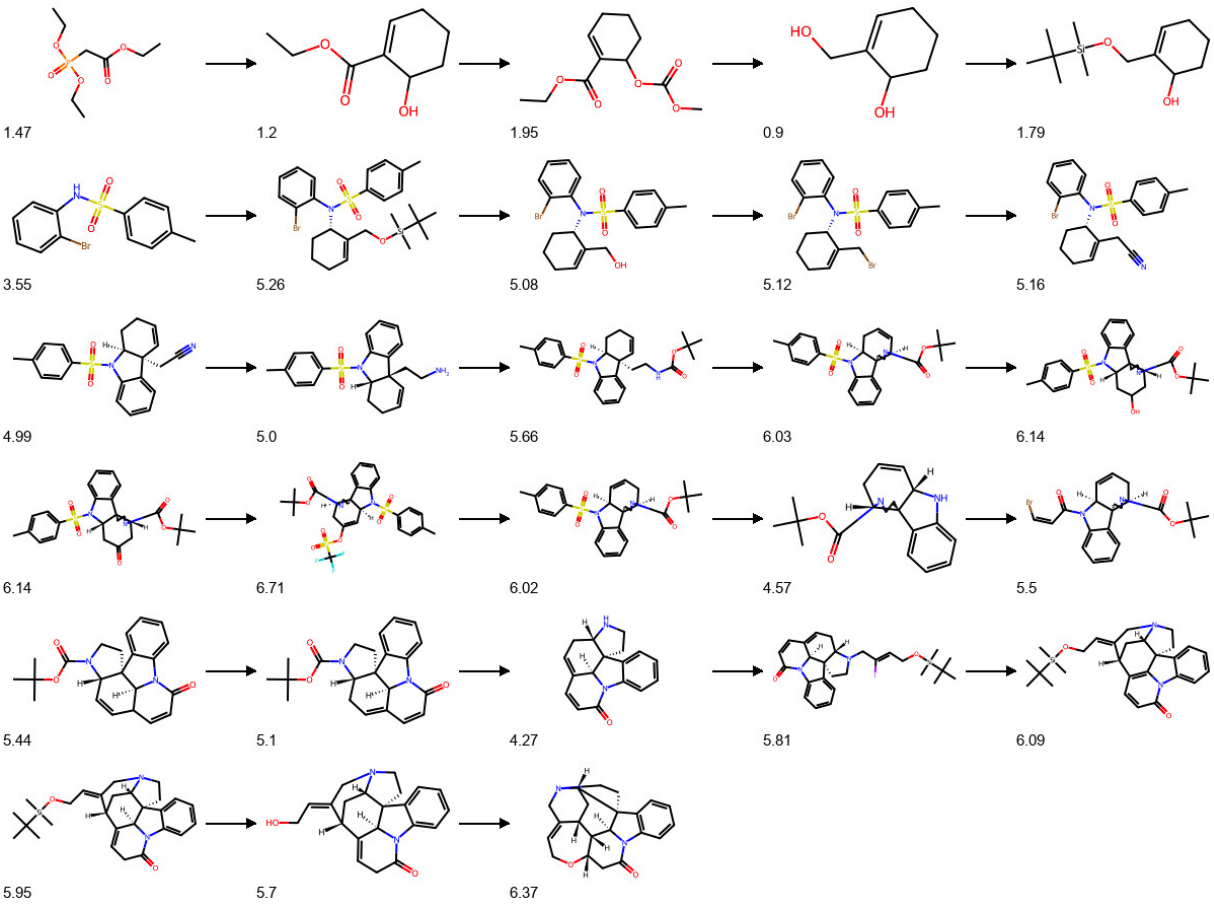
a. Woodward 1954[1]



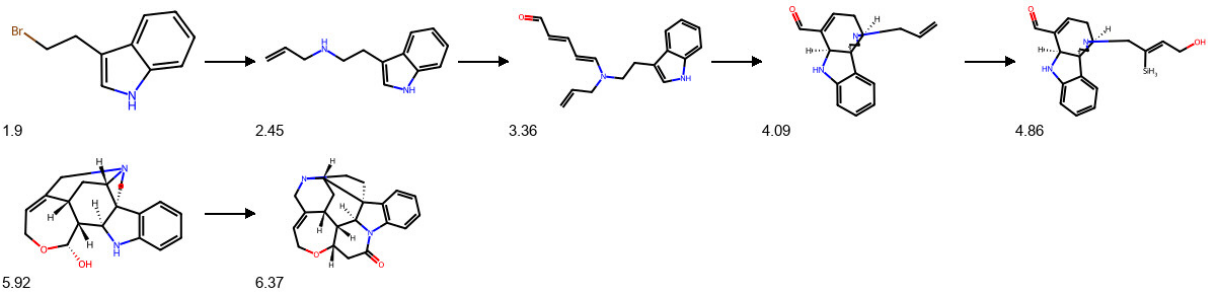
b. Overman 1993[2]



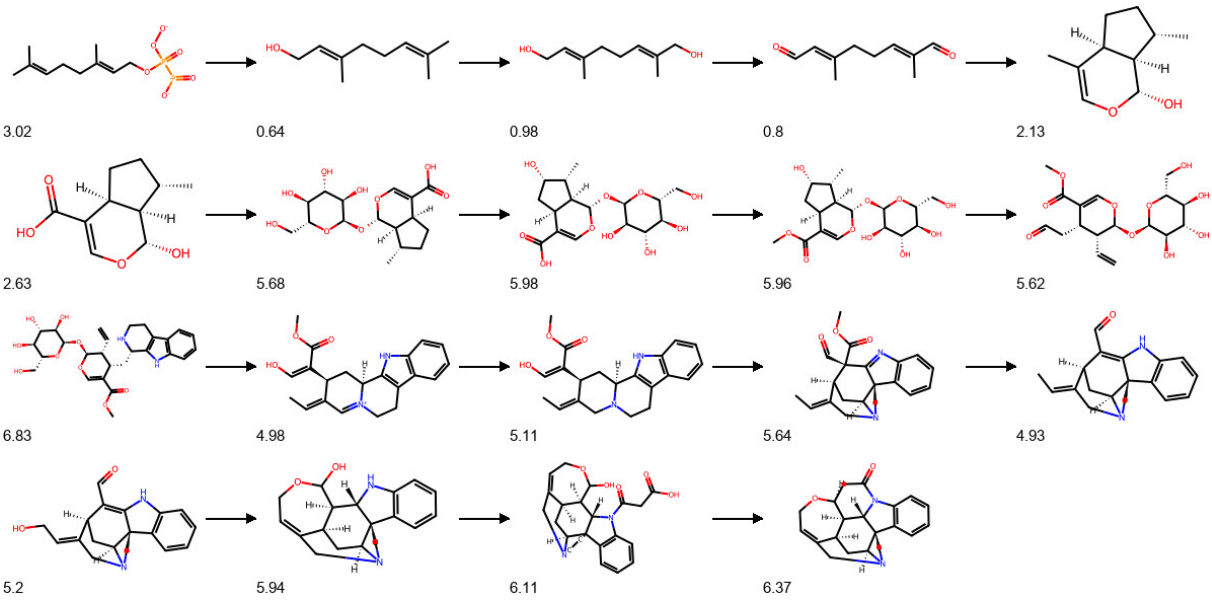
c. Mori 2002[3]


1.47
1.2
1.95
0.9
1.79
3.55
5.26
5.08
5.12
5.16
4.99
5.0
5.66
6.03
6.14
6.14
6.71
6.02
4.57
5.5
5.44
5.1
4.27
5.81
6.09
5.95
5.7
6.37

d. Vanderwal 2011[4]

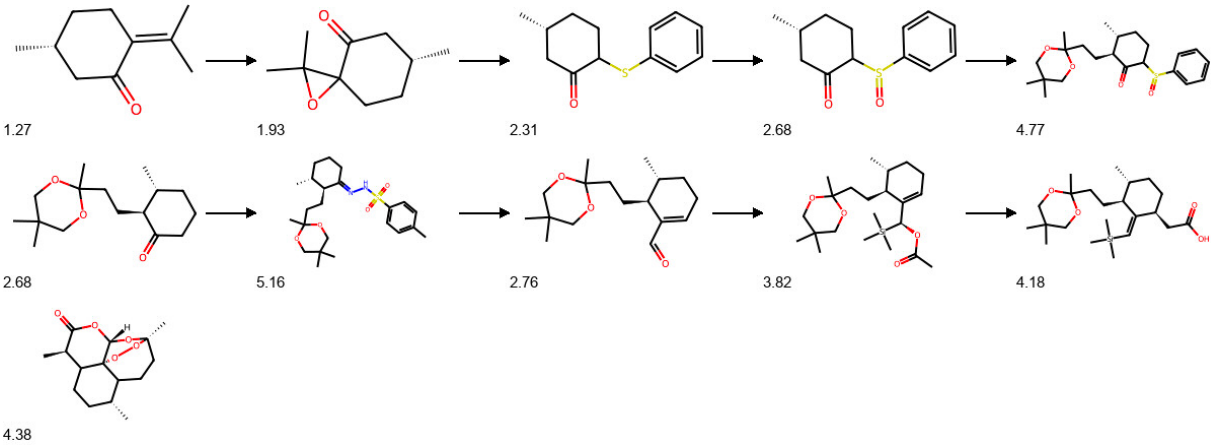


e. Biosynthesis[5]

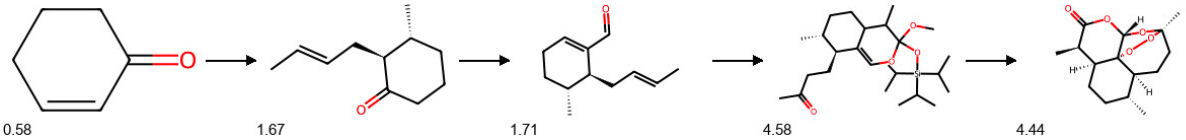


2. Artemisinin

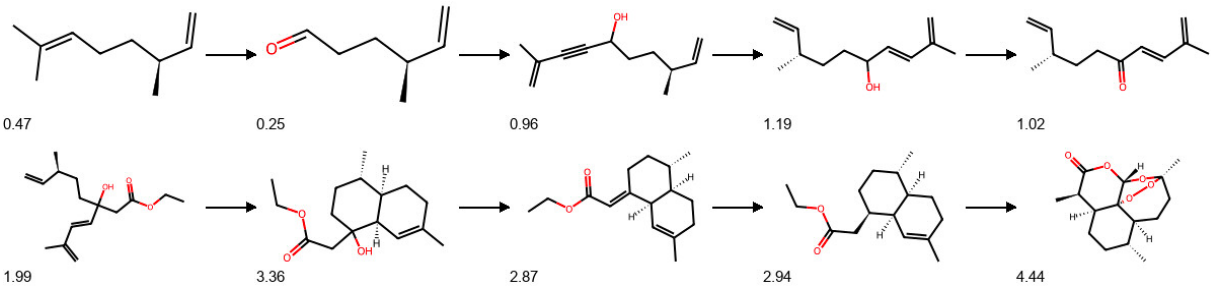
a. Avery 1992[6]



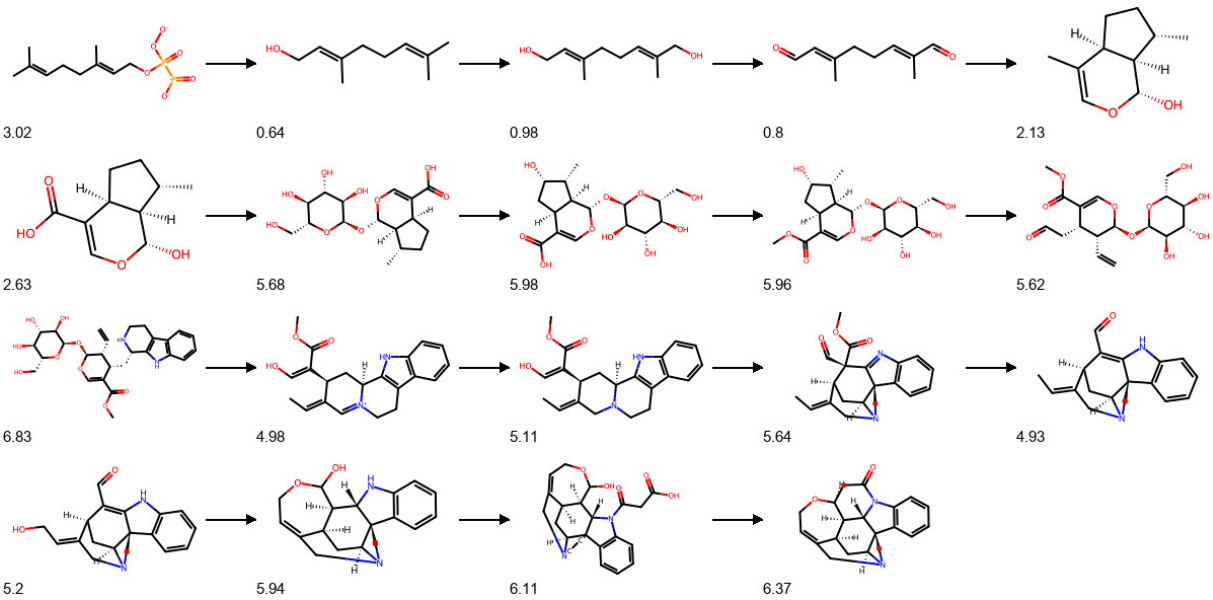
b. Cook 2012[7]



c. Krieger 2018[8]



d. Biosynthesis[9]


3.02
0.64
0.98
0.8
2.13
2.63
5.68
5.98
5.96
5.62
6.83
4.98
5.11
5.64
4.93
5.2
5.94
6.11
6.37

7. Datasets used in the analysis

Several datasets were used during the performed analysis. In particular,

- Figure 1b relies on the sample of PubChem that was obtained according to the description given in the Methods section.
- Results given in Figure 2a-b are the selection of ChEMBL drugs that can be reproduced by using the Datamol framework (the dataset itself can be accessed by datamol.data.chembl_drugs(as_df=True)). Figure 2c involves the analysis of common benchmarking datasets. In particular, QM9, Tox21, and HIV datasets were considered. The latter two were accessed through the therapeutics data commons API.[10]
- Reaction Atlas in Figure 3 is constructed based on the Schneider 50k[11] dataset and is processed according to the methodology developed by Schwaller[12] et al.
- Figure 4a relies on the collection of FDA-approved small molecule drugs, which are publicly available. Additionally, the datasets collected by Ross et al. was used for the quantitative analysis presented in Figure 4b-d. This involves both measurements (that were received as described in corresponding publication[13])as well as docked structures.
- Results in Figure 5 involve the synthetic routes of two natural products that were manually drawn and parsed by the authors of this work.

References

(1) Woodward, R. B.; Cava, M. P.; Ollis, W. D.; Hunger, A.; Daeniker, H. U.; Schenker, K. THE TOTAL SYNTHESIS OF STRYCHNINE. *J Am Chem Soc* **1954**, *76* (18), 4749–4751. https://doi.org/10.1021/ja01647a088.

(2) Knight, S. D.; Overman, L. E.; Pairaudeau, G. Asymmetric Total Syntheses of (-)- and (+)-Strychnine and the Wieland-Gumlich Aldehyde. *J Am Chem Soc* **1995**, *117* (21), 5776–5788. https://doi.org/10.1021/ja00126a017.

(3) Nakanishi, M.; Mori, M. Total Synthesis of (−)-Strychnine. *Angewandte Chemie International Edition* **2002**, *41* (11), 1934. https://doi.org/10.1002/1521-3773(20020603)41:11<1934::AID-ANIE1934>3.0.CO;2-F.

(4) Martin, D. B. C.; Vanderwal, C. D. A Synthesis of Strychnine by a Longest Linear Sequence of Six Steps. *Chem Sci* **2011**, *2* (4), 649. https://doi.org/10.1039/c1sc00009h.

(5) Hong, B.; Grzech, D.; Caputi, L.; Sonawane, P.; López, C. E. R.; Kamileen, M. O.; Hernández Lozada, N. J.; Grabe, V.; O'Connor, S. E. Biosynthesis of Strychnine. *Nature* **2022**, *607* (7919), 617–622. https://doi.org/10.1038/s41586-022-04950-4.

(6) Avery, M. A.; Chong, W. K. M.; Jennings-White, C. Stereoselective Total Synthesis of (+)-Artemisinin, the Antimalarial Constituent of Artemisia Annua L. *J Am Chem Soc* **1992**, *114* (3), 974–979. https://doi.org/10.1021/ja00029a028.

(7) Zhu, C.; Cook, S. P. A Concise Synthesis of (+)-Artemisinin. *J Am Chem Soc* **2012**, *134* (33), 13577–13579. https://doi.org/10.1021/ja3061479.

(8) Krieger, J.; Smeilus, T.; Kaiser, M.; Seo, E.; Efferth, T.; Giannis, A. Total Synthesis and Biological Investigation of (−) Artemisinin: The Antimalarial Activity of

Artemisinin Is Not Stereospecific. *Angewandte Chemie International Edition* **2018**, *57* (27), 8293–8296. https://doi.org/10.1002/anie.201802015.
(9) Ikram, N. K. B. K.; Simonsen, H. T. A Review of Biotechnological Artemisinin Production in Plants. *Front Plant Sci* **2017**, *8*. https://doi.org/10.3389/fpls.2017.01966.
(10) Huang, K.; Fu, T.; Gao, W.; Zhao, Y.; Roohani, Y.; Leskovec, J.; Coley, C. W.; Xiao, C.; Sun, J.; Zitnik, M. Artificial Intelligence Foundation for Therapeutic Science. *Nat Chem Biol* **2022**, *18* (10), 1033–1036. https://doi.org/10.1038/s41589-022-01131-2.
(11) Schneider, N.; Lowe, D. M.; Sayle, R. A.; Landrum, G. A. Development of a Novel Fingerprint for Chemical Reactions and Its Application to Large-Scale Reaction Classification and Similarity. *J Chem Inf Model* **2015**, *55* (1), 39–53. https://doi.org/10.1021/ci5006614.
(12) Schwaller, P.; Probst, D.; Vaucher, A. C.; Nair, V. H.; Kreutter, D.; Laino, T.; Reymond, J. L. Mapping the Space of Chemical Reactions Using Attention-Based Neural Networks. *Nat Mach Intell* **2021**, *3* (2), 144–152. https://doi.org/10.1038/s42256-020-00284-w.
(13) Shi, L.; Pan, H.; Liu, Z.; Xie, J.; Han, W. Roles of PFKFB3 in Cancer. *Signal Transduct Target Ther* **2017**, *2* (1), 17044. https://doi.org/10.1038/sigtrans.2017.44.