Data in Brief 54 (2024) 110417

Contents lists available at ScienceDirect

Data in Brief

journal homepage: www.elsevier.com/locate/dib

Data Article

Dataset of drugs, their molecular scaffolds and medical indications with interactive visualization

Georgii Malakhov[a,b], Pavel Pogodin[a,*]

[a] Institute of Biomedical Chemistry, Pogodinskaya Street, 10, 119121, Moscow, Russia
[b] Faculty of Bioengineering and Bioinformatics, Lomonosov Moscow State University, Leninskie Gory, 1-73, 119991, Moscow, Russia

ARTICLE INFO

Article history:
Received 29 December 2023
Revised 8 April 2024
Accepted 9 April 2024
Available online 14 April 2024

Dataset link: scaffolds-of-the-known-drugs (Original data)

Keywords:
Computational medicinal chemistry
Drug diversity
Bemis-Murcko scaffolds
Interactive visualization
Pharmacological promiscuity

ABSTRACT

Bemis-Murcko scaffolding [1] is a powerful tool for compound clustering and subsequent analysis. Here, using ChEMBL database [2] and RDKit library [3], we have compiled the dataset of known small molecule drugs, their molecular scaffolds and associated medical indications augmented with the interactive interface. We present these data, which can be used by medicinal chemists to find most promising scaffolds for their tasks using an interactive visualization that can help to evaluate both the diversity of known drugs and pharmacological promiscuity of each particular scaffold visually. Our scripts, that are freely available, can help to carry out such scaffold-based analysis and to visualize a compound library in a similar way.

* Corresponding author.
E-mail address: pogodinpv@ibmc.msk.ru (P. Pogodin).

https://doi.org/10.1016/j.dib.2024.110417

Specifications Table

Subject	Drug Discovery
Specific subject area	Analysis of known drugs, clustering by Bemis-Murcko scaffolding, chemical structure visualization, analysis of known drugs' indication and their relation to chemical structure.
Data format	Analyzed, Filtered, Visualised.
Type of data	TSV Tables, JSON, Images
Data collection	Data on approved organic small molecule drugs administered orally or parenterally, excluding any pro-drugs, were extracted from the MySQL ChEMBL of version 33 [2] using PyMySQL Python package. Scaffolds of the extracted drugs and their images were obtained using the RDKit Python package [3]. The final image was composed using Inkscape graphics editor.
Data source location	Institute of Biomedical Chemistry, Pogodinskaya Street, 10 119,121, Moscow, Russia
Data accessibility	Repository name: GitHub DOI: 10.5281/zenodo.10912740 Direct URL to data: https://github.com/RSF-23–73–01058/scaffolds-of-the-known-drugs https://rsf-23–73–01058.github.io/scaffolds-of-the-known-drugs/ Instructions for accessing these data: The data are available through public GitHub repository:https://github.com/RSF-23–73–01058/scaffolds-of-the-known-drugs Also, the data could be accessed and analyzed using interactive interface: https://rsf-23–73–01058.github.io/scaffolds-of-the-known-drugs/

1. Value of the Data

- These data can help to understand both pharmacological promiscuity of the drugs associated with a particular scaffold and diversity of structures of drugs associated with a particular indication. Also, the data can help to perceive the diversity of known drugs in general.
- Computational medicinal chemists can use the data to find the most promising structural scaffolds for their tasks. Pharmacologists could use these data to illustrate the diversity of drugs that are associated with particular medical indications.
- We provide filtered data on known drugs that can be used for a different analysis of the space of by far known drugs. Also, the data on scaffolds of known drugs can be used in computational drug discovery. Last but not least, we provide an example of using an image as an interactive element of the user experience design, so other researchers can depict the diversity of a compound library using a similar approach.

2. Background

The concept of Bemis-Murcko scaffolds [1] has firmly established itself among the cheminformatics tools [4]. Nowadays, it is used for many tasks such as assessing the diversity of chemical libraries [5] and search and generation of chemical structures with the desired properties and certain degree of novelty [6,7]. This motivated us to create a dataset containing scaffolds of FDA approved drugs associated with their medical indications enabling straightforward and reproducible analysis in this field. Furthermore, we were interested to use such dataset as a small, but meaningful example of application of interactive "cloud" chemical diversity visualization [8,9] as an element of database's user interface (https://rsf-23–73–01058.github.io/scaffolds-of-the-known-drugs/).

3. Data Description

TSV table "drugs.tsv" contains the main body of the data presented in this paper. This file contains the following fields: "Parent Compound" which contains the ChEMBL ID of parent compound of one or several drugs, "Preferred Name" which contains the preferred name of the parent compound, "SMILES" which contains canonical SMILES of the parent compound, "Non-generic scaffold SMILES" and "Generic Scaffold SMILES" which contain canonical SMILES of the non-generic scaffold and the generic scaffold of the parent compound accordingly, "Non-generic scaffold ID" and "Generic Scaffold ID" which contain IDs of the non-generic scaffold and the generic scaffold of the parent compound accordingly and "MeSH Indications" which contains MeSH (Medical Subject Headings) terms for the diseases that are associated with the indications that the drugs associated with the parent compound are approved for separated by "|" symbol.

In total the file contains 1155 records, each of which corresponds to a unique parent compound. All of the parent compounds in the file correspond to 2707 drugs. The file contains 820 indications, 521 generic and 788 non-generic scaffolds. The distribution of the number of drugs associated with a particular scaffold is given in Fig. 1.

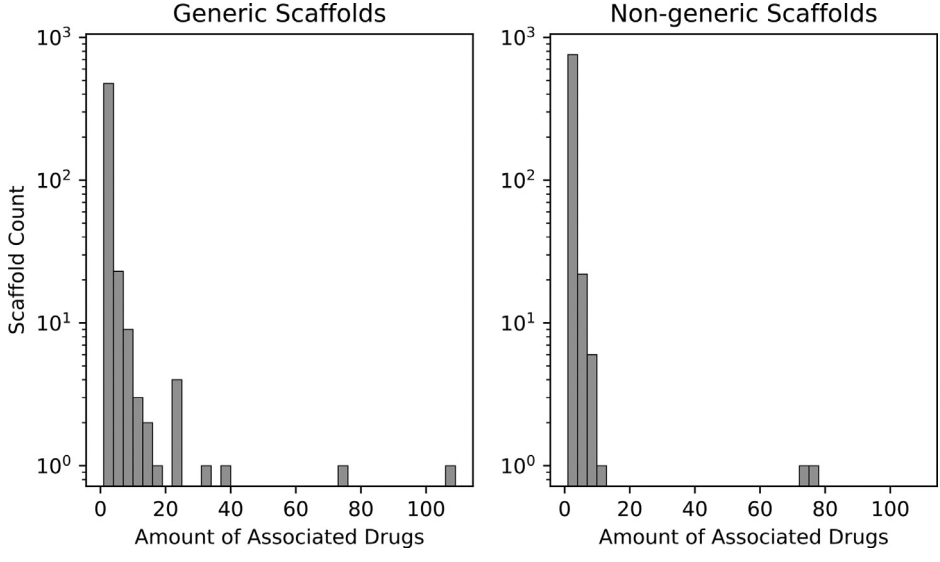

Fig. 1. Distribution of amounts of drugs associated with generic and non-generic scaffolds. Notice that the y-axis scale is logarithmic.

The difference between generic and non-generic scaffolds is presented in Fig. 2. These types of scaffolds were originally presented in the work of Bemis and Murcko [1]. One contains only connectivity of an initial molecule (generic), while the other contains information about each individual atom and bond too (non-generic).

File “drugs.json” contains the same information as “drugs.tsv” but without data on non-generic scaffolds. This format may be better suited for automated processing of the extracted data. The file contains 1155 records that are lists of values of the TSV file’s fields “Parent Compound”, “Preferred Name”, “SMILES”, “Generic Scaffold ID”, “Generic Scaffold SMILES” and “MeSH Indications” in that order. Notice, that MeSH indications in the "drugs.json" are divided by the “ | ” in comparison to the “drugs.tsv” where MeSH indications are divided by the “|” symbol.

The results of clustering and scoring of generic scaffolds are provided as a TSV table named “generic_scaffolds.tsv”. This file contains such fields as “Scaffold ID” which contains the previously given ID of a particular generic scaffold, “Scaffold SMILES” which contains canonical

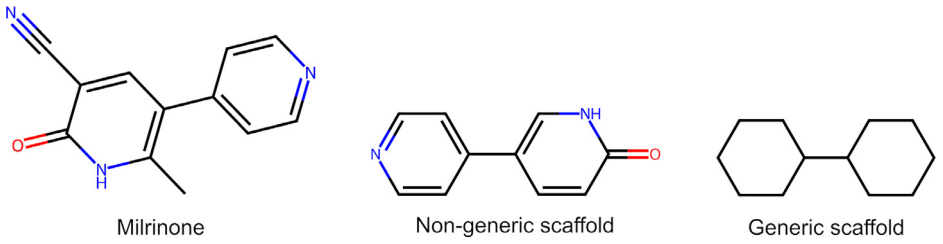

Fig. 2. Chemical structure of Milrinone and its scaffolds. In non-generic scaffolds hybridization and element of each atom along with the order of each bond are presented, whilst in generic scaffolds every atom converted to carbon and every bond – to a single one.

SMILES of the scaffold, "Cluster ID" which contains the ID of the cluster the scaffold belongs to and "Score" which contains scaffold's score from one to ten that correspond to scaffold's frequency of occurrence among the parent compounds of the extracted drugs. In total the file contains 521 records, each of which corresponds to a particular generic scaffold.

The folder "scaffold_svg" that is compressed into the "scaffold_svg.zip" contains folders named from 1 to 94, each of which contains all of the SVG images of scaffolds grouped into a cluster with an ID corresponding to a name of a folder. Also, the folder "scaffold_images" contains a folder named "png" which contains images of all the generic scaffolds in PNG format. Each image either in SVG or in PNG format is named so its name contains the depicted scaffold's cluster ID, its score rounded to four digits without a decimal point and its own scaffold ID in that order separated by underscore symbols.

The file "scaffolds.svg" contains the depictions of all the 521 generic scaffolds in vector format. The size of each scaffold is proportional to its frequency among the extracted parent compounds. The scaffolds are grouped according to the results of the clustering (similar generic scaffolds are more likely to be placed together). To each entity a corresponding ID is assigned. This along with the innate SVG format features allows one to use this file as the template for the interactive applications aiming to navigate the chemical space of known drugs. The file "index.html" integrates the content of the files "scaffolds.svg" and "drug.json" into one interactive page accessible via GitHub Pages (https://rsf-23–73–01058.github.io/scaffolds-of-the-known-drugs/). Users can select drug scaffolds from the "scaffolds.svg" by clicking on their graphical representation and fetch associated information from the "drug.json" in the form of a table, which can be copied or downloaded as .xlsx or .tsv files. And vice versa, users can browse table created on the basis of the "drug.json", select drug or indication which interest them and highlight the corresponding scaffolds from the "scaffolds.svg" on the interactive image.

The file "LICENSE.md" contains the Creative Commons Attribution-Share Alike 3.0 Unported License (https://creativecommons.org/licenses/by-sa/3.0/). This is the license, under which ChEMBL data are provided. This license allows adaptation and redistribution of the data on condition that appropriate attribution is given and that the redistribution is under the same license. Thus, this license regulates the further usage of the data described in this paper.

4. Experimental Design, Materials and Methods

ChEMBL database was chosen as the main data source. ChEMBL is a large and open-access curated database, which contains data on biological activities and chemical structures of approved drugs and investigational compounds [2]. We extracted drug related data from ChEMBL database (version 33), which contains information on about fourteen thousand drugs. To extract the data from ChEMBL we used the MySQL version of the database and PyMySQL Python package to access the database using Python scripts.

Using Python script “get_drug_data.py” we extracted 1155 parent compounds of approved organic small molecule drugs administered orally or parenterally excluding prodrugs, their canonical SMILES, preferred names and indications the drugs are approved for.

Firstly, the script takes data from the table called “DRUG_INDICATION”. If the drug is approved for the observed indication, the script stores this indication and the value of “MOLREGNO” field to access other tables associated with the drug. Then it takes the drug’s parent compound’s “MOLREGNO” from the table “MOLECULE_HIERARCHY”, further analysis is referred to parent compound only. The script takes its ChEMBL ID and preferred name from the table “MOLECULE_DICTIONARY” and stores it only if the parent compound is not a pro-drug, it’s organic and it’s administered orally or parenterally. Then it checks if ChEMBL contains canonical SMILES for the parent compound of the drug, and only if it does, script adds the SMILES to previously obtained information on the parent compound, otherwise it will discard the information on that drug. This test aims to ensure that the parent compound of the drug is a small molecule drug. If all tests are passed by the parent compound, we end up storing indications the drugs associated with the parent compound are approved for, the parent compound’s preferred name and its canonical SMILES associated with the ChEMBL ID of the parent compound. We extracted indications of the drugs as medical subject heading (MeSH) terms of diseases corresponding to indications.

We used the RDKit Python package of version 2023.3.2 [3] to process the chemical structures of the drugs and obtain their Bemis-Murcko scaffolds. The RDKit package is an open-source toolkit for chemoinformatics. Scaffolds of the extracted parent compounds of the drugs were obtained, enumerated and given IDs using “get_drug_data.py”. The JSON version of the table was obtained using Python script “make_json.py”.

Then the extracted generic scaffolds were clustered using the algorithm that was first introduced by Butina et al. [10]. We used the RDKit implementation of this algorithm, scaffolds were clustered using Tanimoto distance. Clusters were enumerated. Then each generic scaffold was given scores from one to ten based on the frequency of occurrence among the extracted drugs. Scoring formula is presented below.

$$score(scaffold) = \frac{9\left(p(scaffold) - \max\limits_{scaffolds} p(scaffold)\right)}{\max\limits_{scaffolds} p(scaffold) - \min\limits_{scaffolds} p(scaffold)} + 10 \quad (1)$$

Clustering and scoring of the extracted generic scaffolds were carried out using Python script "clustering_and_scaling.py".

Each scaffold was drawn in the SVG format using Python script “draw_scaffolds.py”. Each drawing was given its scaffold ID as a value of “id” attribute and its scaffold ID as a value of “cluster” attribute. Values of “class” attribute of the SVG drawings were set as “scaffold”. The scaling factor of each SVG drawing was obtained by multiplication of the scaffold’s score and default scaling factor of the SVG drawer implemented in the RDKit package. Then the final image was constructed by manually placing the scaffold drawings together using the open-source SVG editor Inkscape (https://inkscape.org). Then the final SVG image was checked whether it contains all the scaffolds using Python script “check_svg.py”. Using the Beautiful Soup 4 Python package we collected “id” attribute values from all the “g” tags of the SVG image. These IDs correspond to scaffold IDs, so we checked if the list of IDs from the image contains every ID from one to 521 and if length of the list is 521. The image passed this test, so we consider that it contains all the generic scaffolds extracted.

The interactive interface to the dataset described in this paper was created as follows: Simple HTML page, which includes “scaffolds.svg” (written directly to the HTML document using the <svg></svg> tag) and “drug.json” (as JavaScript variable) was created. Data from “drug.json” were rendered as the table with searchable fields using the DataTables plug-in (https://datatables.net/). The ability to visually highlight, select and deselect scaffolds’ depictions was added using Cascading Style Sheets (CSS). The synchronization of the interactive table’s content and the list of selected scaffolds’ depictions was achieved using several jQuery-aided

(https://jquery.com/) functions listening to the events occurring on the page due to the users' interaction with the interactive content. Overall styling of the page was achieved using the Bootstrap framework (https://getbootstrap.com/).

Limitations

Firstly, we used ChEMBL of version 33, so we analyzed only drugs that are presented there. Secondly, known drugs are diverse, so many of the scaffolds are associated with only one parent compound, and there are more such scaffolds among non-generic scaffolds than among generic scaffolds. That's why to provide more indicative visualization, we ended up using only generic scaffolds for clustering and visualizing. Also, drugs are never studied for every possible indication, so the fact that the drug is not approved for a particular indication doesn't necessarily mean that the drug is not effective in this context. Also, prodrugs were not included in this dataset, thus, some drug classess (antivirals, for example) may be underrepresented.

Ethics Statement

We are acknowledged with the ethical requirements and we can claim that we provide only a secondary dataset, this research doesn't involve any human subjects, any animal experiments, or any data collected from social media.

Data Availability

scaffolds-of-the-known-drugs (Original data) (GitHub).

CRediT Author Statement

Georgii Malakhov: Conceptualization, Methodology, Software, Formal analysis, Data curation, Writing – original draft, Writing – review & editing, Visualization; **Pavel Pogodin:** Conceptualization, Methodology, Software, Writing – review & editing, Visualization, Supervision, Project administration, Funding acquisition.

Acknowledgements

The study was supported by a Russian Science Foundation, grant № 23–73–01058 https://rscf.ru/en/project/23–73–01058/.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] G.W. Bemis, M.A. Murcko, The properties of known drugs. 1. Molecular frameworks, J. Med. Chem. 39 (1996) 2887–2893, doi:10.1021/jm9602928.

[2] D. Mendez, A. Gaulton, A.P. Bento, J. Chambers, M. De Veij, E. Félix, M.P. Magariños, J.F. Mosquera, P. Mutowo, M. Nowotka, M. Gordillo-Marañón, F. Hunter, L. Junco, G. Mugumbate, M. Rodriguez-Lopez, F. Atkinson, N. Bosc, C.J. Radoux, A. Segura-Cabrera, A. Hersey, A.R. Leach, ChEMBL: towards direct deposition of bioassay data, Nucl. Acids Res. 47 (2019) D930–D940, doi:10.1093/nar/gky1075.

[3] rdkit/rdkit: 2023_09_1 (2023) Release Beta, https://doi.org/10.5281/zenodo.8413907
[4] Y. Hu, D. Stumpfe, J. Bajorath, Lessons learned from molecular scaffold analysis, J. Chem. Inf. Model 51 (2011) 1742–1753, doi:10.1021/ci200179y.
[5] M. González-Medina, F.D. Prieto-Martínez, J.R. Owen, J.L. Medina-Franco, Consensus Diversity Plots: a global diversity analysis of chemical libraries, J. Cheminform 8 (2016) 63, doi:10.1186/s13321-016-0176-9.
[6] A. Schuffenhauer, Computational methods for scaffold hopping, WIREs Comput. Mole. Sci. 2 (2012) 842–867, doi:10.1002/wcms.1106.
[7] X. Liu, K. Ye, H.W.T. van Vlijmen, A.P. IJzerman, G.J.P. van Westen, DrugEx v3: scaffold-constrained drug design with graph transformer-based reinforcement learning, J. Cheminform 15 (2023) 24, doi:10.1186/s13321-023-00694-z.
[8] F. Heimerl, S. Lohmann, S. Lange, T. Ertl, Word Cloud explorer: text analytics based on word clouds, in: 2014 47th Hawaii International Conference on System Sciences. Presented at the 2014 47th Hawaii International Conference on System Sciences, 2014, pp. 1833–1842, doi:10.1109/HICSS.2014.231.
[9] P. Ertl, B. Rohde, The molecule cloud - compact visualization of large collections of molecules, J. Cheminform 4 (2012) 12, doi:10.1186/1758-2946-4-12.
[10] D. Butina, Unsupervised data base clustering based on daylight's fingerprint and tanimoto similarity: a fast and automated way to cluster small and large data sets, J. Chem. Inf. Comput. Sci. 39 (1999) 747–750, doi:10.1021/ci9803381.