A new ChEMBL dataset for the similarity-based target fishing engine FastTargetPred: Annotation of an exhaustive list of linear tetrapeptides

Shivalika Tanwar, Patrick Auberger, Germain Gillet, Mario Dipaola, Katya Tsaioun, Bruno Villoutreix

Data in Brief 42 (2022) 108159

Contents lists available at ScienceDirect

Data in Brief

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

Data Article

A new ChEMBL dataset for the similarity-based target fishing engine FastTargetPred: Annotation of an exhaustive list of linear tetrapeptides

Shivalika Tanwar[a], Patrick Auberger[b,c], Germain Gillet[d], Mario DiPaola[e], Katya Tsaioun[e,f], Bruno O. Villoutreix[a,d,*]

[a] Inserm UMR 1141 NeuroDiderot, Robert-Debré Hospital, Université de Paris, Paris 75019, France
[b] Université Côte d'azur, Nice, France
[c] Inserm U1065, C3M, Team 2, Nice, France
[d] Center de Recherche en Cancérologie de Lyon, U1052 INSERM, UMR CNRS 5286, Université de Lyon, Université Lyon I, Center Léon Bérard, 28 rue Laennec, Lyon 69008, France
[e] Akttyva Therapeutics, Inc., MA, Mansfield, USA
[f] Johns Hopkins Bloomberg School of Public Health, MD, Baltimore, USA

ARTICLE INFO

Article history:
Received 9 February 2022
Revised 31 March 2022
Accepted 5 April 2022
Available online 11 April 2022

Dataset link: A new ChEMBL dataset for FastTargetPred (Original data)

Keywords:
Peptide
Virtual screening
Drug discovery
Target prediction

ABSTRACT

Drug discovery often requires the identification of off-targets as the binding of a compound to targets other than the intended target(s) can be beneficial in some cases or detrimental in other situations (e.g., binding to anti-targets). Such investigations are also of importance during the early stage of a project, for example when the target is not known (e.g., phenotypic screening). Target identification can be performed *in-vitro*, but various *in-silico* methods have also been developed in recent years to facilitate target identification and help generate ideas. FastTargetPred is one such approach, it is a freely available Python/C program that attempts to predict putative macromolecular targets (i.e., target fishing) for a single input small molecule query or an entire compound collection using established chemical similarity search approaches. Indeed, the putative macromolecular target(s) of a small chemical compound can be predicted by identify-

* Corresponding author at: Inserm UMR 1141 NeuroDiderot, Robert-Debré Hospital, Université Paris Cité, Paris 75019, France.
E-mail address: bruno.villoutreix@inserm.fr (B.O. Villoutreix).

https://doi.org/10.1016/j.dib.2022.108159

ing ligands that are known experimentally to bind to some targets and that are structurally similar to the input query chemical compound. Therefore, this type of target fishing approach relies on a large collection of experimentally validated macromolecule-chemical compound binding data. The small chemical compounds can be described as molecular fingerprints encoding their structural characteristics as a vector. The published version of FastTargetPred used ligand-target binding data extracted from the release 25 (2019) of the ChEMBL database. Here we provide a new dataset for FastTargetPred extracted from the last ChEMBL release, namely, at the time of writing, ChEMBL29 (2021). Four fingerprints were computed (ECFP4, ECFP6, MACCS and PL) for the extracted compound dataset (714,780 unique ChEMBL29 compounds while the entire ChEMBL29 database contained about 2.1 million compounds). However, it was not possible to compute fingerprints for 19 molecules because of their unusual chemistry (complex macrocycles). These data files were then prepared so as to be compatible with FastTargetPred requirements. The 714,761 ChEMBL chemical compounds with computed fingerprints hit 6,477 macromolecular targets based on the selected criteria. For these ChEMBL compounds a ChEMBL target ID is reported and these target IDs were matched with the corresponding UniProt IDs. Thus, when available, the UniProt ID is provided, the protein UniProt name, the gene name, the organism as well as annotated involvement in diseases, gene ontology data, and cross-references to the Reactome pathway database. As short peptides can be of interest for drug discovery and chemical biology endeavours, we were interested in attempting to predict putative macromolecular targets for a previously reported exhaustive combination of peptides containing four natural amino acids (i.e., $20 \times 20 \times 20 \times 20 = 160{,}000$ linear tetrapeptides) using FastTargetPred and the presently generated ChEMBL29 dataset. With the parameters used, putative targets are reported for 63,944 unique query peptides. These target predictions are provided in two different searchable files with hyperlinks to the ChEMBL, UniProt and Reactome databases.

Specifications Table

Subject	Drug Discovery
Specific subject area	Macromolecular target predictions for input small chemical compounds, Target Fishing.
Type of data	Text files (CSV, TXT) DataWarrior files
How the data were acquired	Chemical compounds and corresponding macromolecular targets were extracted from the ChEMBL29 database [1]. ChEMBL target IDs were mapped onto the UniProt target IDs [2] so as to collect UniProt protein names, gene names, possible involvements in diseases, gene ontology data [3] and pathway identifiers present in the Reactome knowledgebase [4].

(*continued on next page*)

	160,000 linear tetrapeptides (SMILES strings) were downloaded from [5] and merged into a single file. Putative targets for these peptides were predicted using FastTargetPred [6] and the newly generated ChEMBL29 dataset. The DataWarrior software [7] was used for data-visualization. Chemical file curation and computations of fingerprints were performed with MayaChemTools [8] while file format conversions were carried out with Open Babel [9].
Data format	Raw and processed data: Extracted ChEMBL compounds (canonical SMILES and ChEMBL compound IDs) in TXT format. ChEMBL compound IDs and corresponding ChEMBL target IDs in TXT format (name of the file = chembl29_active_all.tlt). Computed fingerprints in CSV format with the corresponding binary files (to accelerate the similarity search computations) readable by FastTargetPred (four bfp files are provided). The ChEMBL – UniProt mapping file is in CSV format. The predicted targets for the tetrapeptides are reported in two files that were generated with the open source DataWarrior software (dwar files). These files are searchable (e.g., using amino-acid sequence query…), they can be sorted by column values and hyperlinks to the ChEMBL, UniProt and Reactome databases were added when available.
Description of data collection	The SQLite version of the ChEMBL29 database was downloaded. SQLite terminal shell commands were applied to extract an initial data file (this raw data file is provided in the extra_data folder). Canonical SMILES strings were extracted for small molecules associated with a biological assay involving a single protein or a protein complex, and the selected assay type was "binding". Only molecules with experimental potency/affinity/activity data (pChEMBL_value) corresponding to 20 micro-molar or less were selected. This initial raw file contained 1,412,822 compounds. Additional filtering involved selecting molecules with a ChEMBL confidence_score of 6 or above, manual curation of mixtures and removal of compounds with obvious errors. Salts were removed with MayaChemTools. The resulting filtered file containing 714,780 unique ChEMBL29 bioactive compounds is provided (available in the extra_data folder). The downloaded tetrapeptide files were merged and used as input for FastTargetPred. Predictions could be performed for some query peptides; two different annotated files with the prediction are reported (located in the "tetrapeptides" folder).
Data source location	• Inserm U1141, Hospital Robert Debre, 75019 • City: Paris • Country: France
Data accessibility	The data are freely available on the Zenodo open access platform and accessible via the following link: https://zenodo.org/record/5751499 And here: Mendeley Data: https://data.mendeley.com/datasets/9t5zdgs3s2/1
Related research article	L. Chaput, V. Guillaume, N. Singh, B. Deprez, B.O. Villoutreix, FastTargetPred: a program enabling the fast prediction of putative protein targets for input chemical databases, Bioinformatics. 36 (2020) 4225–4226. 10.1093/bioinformatics/btaa494

Value of the Data

- This dataset should be valuable to research groups interested in carrying out macromolecular target predictions or in investigating bioactive molecules.
- These data could be integrated into other tools and manipulated by different types of algorithms.
- These data used together with FastTargetPred or related tools could be useful for understanding the polypharmacology and safety profile of existing or virtual small molecules.
- Data generated on tetrapeptides could be useful as a starting point for drug discovery projects or chemical biology studies.

- Users can enrich these data by performing *in-vitro* screening campaigns or by mining other repositories.

1. Data Description

To be able to carry out target prediction via ligand-based similarity search, a large collection of experimentally validated ligand-target pairs must be generated [10–12]. We extracted such annotated data from the last version (at the time of writing) of the ChEMBL database (release 29) [1]. A first initial file (raw data) was thus generated containing 1,412,822 compounds (file located in the extra_data folder). Further filtering of the compounds was performed to remove annotation errors, molecules that did not have a high ChEMBL confidence_score, and a pChEMBL value. Salts and duplicates were also removed using Unix text mining commands, Open Babel [9], and MayaChemTools [8] with visualization in DataWarrior [7]. These steps led to the generation of a filtered file with 714,780 compounds in SMILES format with the associated ChEMBL compound IDs (file located in the folder named "extra_data"). Four fingerprints were computed with MayaChemTools for each compound. These four files are in CSV format and loaded in a single directory located in the "extra_data" folder. The fingerprints for 19 molecules could not be computed and as such are not present in the CSV files. A folder specific for FastTargetPred was generated and named "dbchembl29". This one contains different files (4 bfp, 1 tlt and 1 CSV files) as required by the software. For the four fingerprints files, the corresponding binary files were generated to accelerate the similarity search computations. These four files have an extension bfp. Then, a text file with extension tlt contains the ChEMBL compound IDs in the first columns and in the following columns, the corresponding ChEMBL target IDs. Finally, ChEMBL target IDs were matched with the UniProt protein IDs. Additional information were extracted via the UniProt server such as gene name, gene ontology data, possible involvement in diseases for each target (if known) and Reactome pathway identifiers. This file is required and is named uniprot_database_ChEMBL.csv. For a user, upon installation of FastTargetPred (https://github.com/ludovicchaput/FastTargetPred) and MayaChemTools (http://www.mayachemtools.org/), it will be necessary to go to the main FastTargetPred folder, and then launch a basic command to search targets for a ligand in SDF format, for instance: python3 FastTargetPred.py rivaroxaban.sdf (the file containing the ligand rivaroxaban is provided in the extra_data folder). This command uses the previously generated data for FastTargetPred extracted from a previous release of ChEMBL (release 25) (the ChEMBL25 files are in the "db" folder). Now, the user can download the newly generated files extracted from ChEMBL (release 29) and put the folder "dbchembl29" at the same level than the "db" folder and run the following command: python3 FastTargetPred.py rivaroxaban.sdf -fp ECFP4 -tc 0.7 -db dbchembl29/chembl29_active. This will allow the package to search for the new data present in the "dbchembl29" folder using all the files that start with chembl29_active and the CSV file. In this example the Tanimoto coefficient is set to 0.7 (i.e., with such a value and with this fingerprint one is looking for molecules, and their corresponding targets, that should be very similar to rivaroxaban as well as rivaroxaban itself as it is present in the ChEMBL29 dataset). Note that several fingerprints can be used and that a consensus score can be generated (please see the FastTargetPred user guide presents in the package).

Applying ligand-based similarity search to identify targets for peptides (e.g., short peptides, chemically modified peptides that are not easy to investigate using traditional amino acid sequence search, etc.) is still underexploited. We decided to apply the newly extracted data and FastTargetPred to 160,000 linear tetrapeptides previously released in SMILES format (available here: https://data.mendeley.com/datasets/z8zh5rpthg/1) [5]. Putative targets were found for 63,944 query peptides using ECFP4 fingerprints and a Tanimoto coefficient of 0.6 (that we evaluated to be reasonable for this type of investigation with FastTargetPred [6]). Clearly, additional investigations will be needed but for now, some of these predictions can be used to generate novel hypotheses. All the results are reported in two DataWarrior files present in the "tetrapeptides" folder. In these files, the query peptide in SMILES format (and the corresponding amino

acids) are reported and so are the ChEMBL compounds found to be (relatively) similar to the queries, the Tanimoto coefficient, the putative targets, the possible links to diseases, the UniProt IDs and the Reactome identifiers. Images of the compounds are present. These files can be sorted by columns, and search by keywords. Hyperlinks to the ChEMBL, UniProt and Reactome databases have been inserted. Obviously, users can apply other types of fingerprints and different Tanimoto coefficients or use consensus scoring via the selection of several fingerprints.

We compared the chemical space covered by the 714,780 ChEMBL29 bioactive compounds (orange spheres), the 160,000 tetrapeptides (blue spheres) and approved drugs (magenta spheres) extracted from DrugBank 5.0 [13] using principal component analysis (PCA) based on six important physicochemical properties in the field of drug discovery (computed within DataWarrior): molecular weight, hydrogen bond donors, hydrogen bond acceptors, topological polar surface area, logP (the octanol/water partition coefficient, i.e., notion of lipophilicity) and number of rotatable bonds (Fig. 1). The bioactive ChEMBL29 molecules covers most of the property space yet, some tetrapeptides occupy a different region and could be interesting to investigate further experimentally as they could hit novel targets. As somewhat expected, the approved drugs are essentially contained within the ChEMBL space (i.e., ChEMBL compounds can be considered as non-optimized chemical probes with some having the right profile to become a drug).

2. Experimental Design, Materials and Methods

The SQLite version of the ChEMBL29 database was downloaded from the ChEMBL website. SQLite standard terminal shell commands were applied to collect the molecules. The canonical SMILES strings of the compounds were extracted for small molecules with an associated biological assay involving a single protein or a protein complex and an assay type named "binding" (assays measuring binding of compound to a molecular target, e.g. Ki, IC50). Thus small molecules with functional assays measuring biological effects without associated targets were not considered; only bioactivities with pChEMBL values were chosen. This term refers to all the comparable measures of half-maximal responses (e.g., IC_{50}, EC_{50}, Ki, Kd…) on a negative logarithmic scale. The considered activity value for a compound-target pair corresponds to a value of 20 micro-molar or less. Additional filtering steps were then applied to this initial raw data file. A confidence_score was already assigned to each assay-to-target relationships in the ChEMBL database with ranges from 0 (uncurated data entries) to 9, with 9 representing a single macromolecular binding target that has been assigned to a compound. We selected small molecules with a confidence_score running from 6 (homologous protein complex subunits assigned) to 9. Mixtures were curated manually while salts were removed using MayaChemTools. Duplicates were investigated by text mining over the ChEMBL compound IDs and using Open Babel. The resulting filtered file contained 714,780 unique ChEMBL29 compounds acting on 6477 different macromolecular targets. By contrast with the previous ChEMBL dataset (extracted from ChEMBL25, it contained 524,810 unique ChEMBL25 compounds active on 4811 targets) that is presently embarked with FastTargetPred package. Furthermore, we did not remove PAINS (pan-assay interference compounds) molecules because of numerous debates on the topic, as not all PAINS can be problematic, and many PAINS compounds can possess specific activity and can co-crystallize with a target; as such they can be investigated more in depth at a later stage [14–17].

Three types of well-established fingerprints as implemented in MayaChemTools were computed: Extended-Connectivity Fingerprints (ECFP, we selected two highly used methods: ECFP4 (atom radius 2) and ECFP6 (atom radius 3) encoded on 1024 bits), Path Length (PL, enumeration of linear fragments with various length encoded on 1024 bits) and Molecular ACCess System (MACCS, for which each bit indicates the presence or absence of specific atoms and chemical substructures, the 322 bits version was selected). Fingerprints could not be computed for 19 compounds present in the extracted ChEMBL29 dataset because of unusual chemistry (compound IDs: CHEMBL1160519, CHEMBL2012126, CHEMBL2012127, CHEMBL2012128, CHEMBL2313980, CHEMBL2313981, CHEMBL2373427, CHEMBL2373486, CHEMBL2373969, CHEMBL2373970, CHEMBL269742, CHEMBL3354724, CHEMBL3354729, CHEMBL3354735,

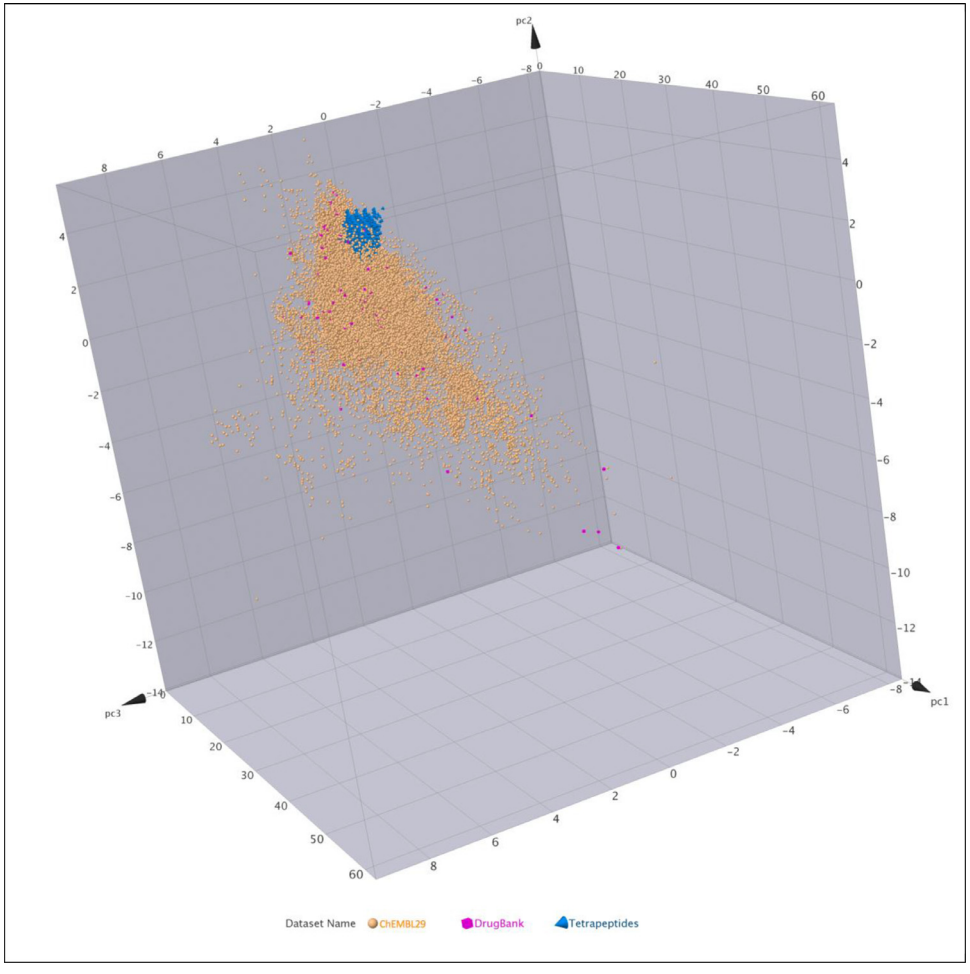

Fig. 1. Chemical space. Comparison of the chemical space covered by the extracted ChEMBL bioactive compounds (orange), tetrapeptides (blue) and approved drugs (magenta) obtained after filtering DrugBank 5.0 (downloaded in December 2021). 2509 approved drugs were initially collected but 287 very small compounds (molecules with less than 10 heavy atoms) were removed (e.g., several compounds have only 1 atom). In addition, 124 other molecules (unusual chemistry and some mixtures) were deleted. Six physicochemical properties were computed with DataWarrior and a PCA plot was generated. The explained variance percentage of the first 3 principal components are PCA1: 81.79%, PCA2: 13.27%, PCA3: 2.52%.

CHEMBL384970, CHEMBL409633, CHEMBL412073, CHEMBL414976, CHEMBL440623), leading to a final number of 714,761 small molecules with associated fingerprints. The Tanimoto coefficient was used as a metric to measure the similarity between a query compound and an annotated compound present in the ChEMBL database [10–12]. This coefficient ranged from 0 (molecules are not similar) to 1 (molecules are identical or very similar). ChEMBL target IDs were matched with UniProt IDs via the UniProt website and the IDs of about 100 targets had to be manually fixed. When available, protein UniProt names, the gene names, the organism, the involvement in diseases of the target, gene ontology data and cross-references to the Reactome pathway database were also compiled. The CSV file containing these annotations is provided (file name: uniprot_database_ChEMBL.csv) as required for the proper functioning of FastTargetPred.

160,000 linear tetrapeptides were previously reported. These peptides are described as SMILES strings and were converted to structure-data file (SDF) using Open Babel. ECFP4 Fingerprints were computed for all these peptides with MayaChemTools. Target prediction for each peptide was performed with FastTargetPred and the newly extracted ChEMBL29 data. A Tanimoto threshold of 0.6 was used to consider two molecules as similar. The target prediction results are reported in two DataWarrior files with information about the query peptides (amino acid composition and SMILES strings), the chemical present in ChEMBL29 predicted as similar with the associated Tanimoto coefficient, the ChEMBL compound IDs (with hyperlinks), the SMILES strings, the target UniProt names if known (with hyperlinks), the ChEMBL target IDs (with hyperlinks), the gene name, the organism, the potential role in disease, gene ontology data and Reactome identifiers (with hyperlinks). Images of the query compounds and of the compounds found to be similar are also present in these two files. These table files can be sorted and searched using keywords. Many additional molecular descriptors can then be computed within DataWarrior as needed by the users.

Ethics Statements

These are secondary datasets that did not involve any human or animal testing.

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.

Data Availability

A new ChEMBL dataset for FastTargetPred (Original data) (Mendeley Data).

CRediT Author Statement

Shivalika Tanwar: Data curation, Formal analysis, Writing – original draft; **Patrick Auberger:** Writing – review & editing; **Germain Gillet:** Writing – review & editing; **Mario DiPaola:** Writing – review & editing; **Katya Tsaioun:** Writing – review & editing; **Bruno O. Villoutreix:** Conceptualization, Data curation, Formal analysis, Supervision, Writing – review & editing.

Acknowledgments

Funding: This publication is partially supported by National Science Foundation Award No. 2136307 and by the PRTK Inca ONCO-BCL and ARC Nrh Awards.

References

[1] 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, Nucleic Acids Res. 47 (D1) (2019) D930–d940, doi:10.1093/nar/gky1075.

[2] UniProt Consortium, UniProt: the universal protein knowledgebase in 2021, Nucleic Acids Res. 49 (D1) (2021) D480–D489, doi:10.1093/nar/gkaa1100.

[3] M. Ashburner, C.A. Ball, J.A. Blake, D. Botstein, H. Butler, J.M. Cherry, A.P. Davis, K. Dolinski, S.S. Dwight, J.T. Eppig, M.A. Harris, D.P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J.C. Matese, J.E. Richardson, M. Ringwald, G.M. Rubin, G. Sherlock, Gene ontology: tool for the unification of biology. The gene ontology consortium, Nat. Genet. 25 (1) (2000) 25–29, doi:10.1038/75556.

[4] M. Gillespie, B. Jassal, R. Stephan, M. Milacic, K. Rothfels, A. Senff-Ribeiro, J. Griss, C. Sevilla, L. Matthews, C. Gong, C. Deng, T. Varusai, E. Ragueneau, Y. Haider, B. May, V. Shamovsky, J. Weiser, T. Brunson, N. Sanati, L. Beckman, X. Shao, A. Fabregat, K. Sidiropoulos, J. Murillo, G. Viteri, J. Cook, S. Shorser, G. Bader, E. Demir, C. Sander, R. Haw, G. Wu, L. Stein, H. Hermjakob, P. D'Eustachio, The reactome pathway knowledgebase 2022, Nucleic Acids Res. 50 (D1) (2022) D687–d692, doi:10.1093/nar/gkab1028.
[5] V.D. Prasasty, E.P. Istyastono, Data of small peptides in SMILES and three-dimensional formats for virtual screening campaigns, Data Brief 27 (2019) 104607, doi:10.1016/j.dib.2019.104607.
[6] L. Chaput, V. Guillaume, N. Singh, B. Deprez, B.O. Villoutreix, FastTargetPred: a program enabling the fast prediction of putative protein targets for input chemical databases, Bioinformatics 36 (14) (2020) 4225–4226, doi:10.1093/bioinformatics/btaa494.
[7] T. Sander, J. Freyss, M. von Korff, C. Rufener, DataWarrior: an open-source program for chemistry aware data visualization and analysis, J. Chem. Inf. Model. 55 (2) (2015) 460–473, doi:10.1021/ci500588j.
[8] M. Sud, MayaChemTools: an open source package for computational drug discovery, J. Chem. Inf. Model. 56 (12) (2016) 2292–2297, doi:10.1021/acs.jcim.6b00505.
[9] N.M. O'Boyle, M. Banck, C.A. James, C. Morley, T. Vandermeersch, G.R. Hutchison, Open babel: an open chemical toolbox, J. Cheminform. 3 (2011) 33, doi:10.1186/1758-2946-3-33.
[10] A. Cereto-Massagué, M.J. Ojeda, C. Valls, M. Mulero, G. Pujadas, S. Garcia-Vallve, Tools for in silico target fishing, Methods 71 (2015) 98–103, doi:10.1016/j.ymeth.2014.09.006.
[11] N. Singh, L. Chaput, B.O. Villoutreix, Virtual screening web servers: designing chemical probes and drug candidates in the cyberspace, Briefings Bioinf. 22 (2) (2021) 1790–1818, doi:10.1093/bib/bbaa034.
[12] N. Mathai, J. Kirchmair, Similarity-based methods and machine learning approaches for target prediction in early drug discovery: performance and scope, Int. J. Mol. Sci. 21 (10) (2020), doi:10.3390/ijms21103585.
[13] 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, M. Wilson, DrugBank 5.0: a major update to the DrugBank database for 2018, Nucleic Acids Res. 46 (D1) (2018) D1074–d1082, doi:10.1093/nar/gkx1037.
[14] J.B. Baell, G.A. Holloway, New substructure filters for removal of pan assay interference compounds (PAINS) from screening libraries and for their exclusion in bioassays, J. Med. Chem. 53 (7) (2010) 2719–2740, doi:10.1021/jm901137j.
[15] S.J. Capuzzi, E.N. Muratov, A. Tropsha, Phantom PAINS: problems with the utility of alerts for pan-assay interference compounds, J. Chem. Inf. Model. 57 (3) (2017) 417–427, doi:10.1021/acs.jcim.6b00465.
[16] D. Lagorce, N. Oliveira, M.A. Miteva, B.O. Villoutreix, Pan-assay interference compounds (PAINS) that may not be too painful for chemical biology projects, Drug Discov. Today 22 (8) (2017) 1131–1133, doi:10.1016/j.drudis.2017.05.017.
[17] E. Gilberg, M. Gütschow, J. Bajorath, X-ray structures of target-ligand complexes containing compounds with assay interference potential, J. Med. Chem. 61 (3) (2018) 1276–1284, doi:10.1021/acs.jmedchem.7b01780.