1

Open Access Databases and Datasets for Computer-Aided Drug Design. A Short List Used in the Molecular Modelling Group of the SIB

Antoine Daina[1], María José Ojeda-Montes[1], Maiia E. Bragina[2], Alessandro Cuozzo[2], Ute F. Röhrig[1], Marta A.S. Perez[1], and Vincent Zoete[1,2]

[1] SIB Swiss Institute of Bioinformatics, Molecular Modeling Group, Quartier UNIL-Sorge, Bâtiment Amphipôle, CH-1015 Lausanne, Switzerland
[2] University of Lausanne, Ludwig Institute for Cancer Research, Department of Oncology UNIL-CHUV, Route de la Corniche 9A, CH-1066 Epalinges, Switzerland

The role of computer-aided drug design (CADD) in modern drug discovery [1–15] is to support its various processes, including hit finding, hit-to-lead, lead optimization, and the activities preluding to preclinical trials, through numerous in silico predictors and filters. These tools have a wide variety of objectives, such as enriching the families of molecules that will be submitted to experimental screening with potentially active compounds, identifying molecules that may be problematic such as toxic moieties or those with nonspecific activities, generating ideas on the chemical modifications to be made to the compounds to increase their affinity for the therapeutic target or to improve their pharmacokinetics [16–19], or finally assisting in the various selection processes aimed at identifying and promoting the most promising molecules. These approaches are generally divided into two main families [20].

Structure-based approaches [8, 21–23] use the three-dimensional structure of the targeted protein, for example, to estimate via the use of a docking software how and how strongly a small molecule will bind to it. Avoiding the necessity to resort solely to an experimental method (*e.g.* X-ray crystallography, NMR, or cryo-electron microscopy) to obtain this information makes it possible to process a large number of molecules very quickly and at a moderate cost. In turn, this information can be used to determine how to modify the chemical structure of a small molecule to optimize rationally the intermolecular interactions with the protein target. It is then possible to select the most promising compounds for experimental validations, creating a cyclic optimization process, thanks to this feedback loop between *in silico* and *in vitro* approaches.

Ligand-based approaches take advantage of already known molecules with certain bioactivities or physicochemical properties, in order to derive the information necessary to predict the bioactivity or properties of other compounds, real or virtual. Indeed, CADD has been a pioneering research area in the development and application of machine learning methods [24–32], with the emergence, as early as the

Open Access Databases and Datasets for Drug Discovery, First Edition.
Edited by Antoine Daina, Michael Przewosny, and Vincent Zoete.

1960s [33], of quantitative structure–activity relationships (QSAR [34]) or quantitative structure–property relationships (QSPR).

To perform these tasks, CADD benefits from numerous databases and datasets of small molecules, bioactivities and biological processes, 3D structures of small compounds and biomacromolecules, or molecular properties – some of which being related to pharmacokinetics or toxicity [13, 35–38]. Created in 1971, the Protein Data Bank (PDB) [39], which stores the three-dimensional structural data of large biological molecules such as proteins and nucleic acids, is a precursor in the field of freely and publicly available databases with possible applications in CADD. Currently managed by the wwPDB [40] organization and its five members, RCSB PDB [41], PDBe [42], PDBj [43], EMDB [44] and BMRB [45], the PDB continues to provide the CADD community with numerous valuable 3D structures of therapeutically relevant proteins in the apo form or in complex with small drug-like molecules, which can be used to nurture structure-based approaches. Several subsets involving such structures have been created over time, for instance, to provide reference sets to benchmark docking software, such as the Astex [46] or the Iridium [47] datasets. For a very long time, ligand-based approaches were generally limited to the use of small datasets, collected on a case-by-case basis during specific drug design projects, thus precluding their application beyond the building of focused models with limited scope. This situation dramatically changed during the 2000s with the rise of large-scale databases created specifically for the benefit of drug discovery in general and CADD in particular. ChEMBL [48, 49] released in 2008 or PubChem [50] in 2004, which collect molecules and their activities in biological assays systematically extracted from medicinal chemistry literature, patent publications, or experimental high-throughput screening programs, are certainly among the forerunners of this trend. Such databases paved the way for CADD approaches addressing, for instance, the prediction of bioactivities on a very large scale, including ligand-based methods. ZINC [51], freely accessible from 2004, is another large-scale database of small molecules, this time prepared especially for virtual screening. This important resource focuses on the compilation and storage of commercially available chemical compounds. DrugBank [52], whose first version dates back to 2006, is an example of a database gathering numerous curated and high-quality information about a group of molecules of biological interest, in this case mainly but not exclusively, approved or developmental drugs. Although smaller than ChEMBL or PubChem for instance, this type of resources, because of the quality, the structure and the practicality of the information provided, also plays an critical role in the development of new CADD techniques and filters, or for more direct applications in virtual screening.

Researchers working in CADD can be considered to have two main activities: one consists in designing, validating, and benchmarking new *in silico* approaches, the other is applying existing tools to support drug descovery projects. The nature of the databases reflects this duality. Some are clearly oriented toward an applicative usage. With virtual screening in mind, this is the case for resources gathering a large amount of commercial or virtual molecules, such as ZINC [51] or GDB-17 [53], whose main purpose is to be used as a source of molecules to feed virtual screening campaigns. At the opposite end of the spectrum, we find molecular sets constructed specifically for benchmarking screening methods, such as DUD-E [54] or DEKOIS [55]. These contain a limited number of compounds, known to be active or inactive on certain protein targets, and carefully chosen to avoid any bias in many molecular properties that would allow a screening software to identify the active ones too easily. Between these two extremes, we can find databases, such as ChEMBL, PubChem, or TCRD/Pharos [56], containing a large number of known bioactive molecules. These generalist databases can not only be used to develop a large range of CADD methods, including screening or reverse screening approaches, such as Similarity Ensemble Approach (SEA) [57, 58] or SwissTargetPrediction [59, 60], but also constitute a source of *real* molecules to be virtually screened.

By definition, the interest for many CADD-related databases lies in their capacity to store a possibly large quantity of molecules, along with useful annotations, and in their efficient diffusion to the public. This was made possible by the development and dissemination of widely accepted specific file formats. The most common file for representing molecules as strings are in SMILES [61, 62] and InChI [63, 64] formats. These one-line formats have the great advantage of using little disk or memory resources, facilitating the storage, and rapid transfer of large numbers of molecules. It should be noted, however, that several SMILES strings can represent the same molecule. This can be problematic and potentially generate redundancy when compounds from different sources are gathered. To avoid this kind of situation, it is possible to produce canonical SMILES by a well-chosen software, which are by definition unique for each molecule, or to use the UniChem [65] database that provides pointers between the molecules of most common databases. Structure-based approaches, such as molecular docking, 3D fingerprinting [66], or pharmacophores [67, 68], require a spatial representation of small molecules. The most frequently employed file definitions, including tridimensional atomic coordinates, are the Structural Data File (SDF), the MDL Mol, and Tripos Mol2 formats. Compounds are often available in such formats in the major small-molecule databases, such as ZINC [51], Chemspider [69], or DrugBank [52], which allow their direct use in 3D-based approaches. Other formats are available to store 3D structures of biomacromolecules, taking advantage of the fact that large biomolecules are based on the repetition of a small number of residues. The PDB and mmCIF [70] formats are among the standards and provided by the wwPDB consortium, and by other major databases of 3D structures of macromolecules, including PDB Redo [71, 72], as well as the SWISS-MODEL [73], MODBASE [74], and AlphaFold [75, 76] repositories of structural models.

To be valuable in the context of CADD, a database should meet several criteria in addition to the nature of its content. These criteria are very close to the findability, accessibility, interoperability, and reuse (FAIR) principles [77].

First, a database must be maintained and made available for the long term, ideally via a persistent URL, so that it can be employed for sustainable projects and developments. Unfortunately, a large fraction of new databases and datasets disappear only a few years after their initial release, due to lack of resources to maintain them or lack of interest. Attwood and colleagues studied the 18-year survival status of 326 databases published before 1997 and found that 62.3% were dead, 14.4% were archived (and not updated), and only 23.3% were still alive under their original identity or after rebranding [78]. This first analysis was independently confirmed by Finkelstein et al. who found that of the 518 original databases published in the journal *Database* between 2009 and 2016, 35% were already no longer accessible in 2020 [79], and by Imker who observed that among the 1727 databases published between 1991 and 2016 in *Nucleic Acids Research*'s "Database Issue," 40% were dead in 2018 [80]. They found that databases with higher citation counts and from researchers with higher h-index within renowned institutions were more likely to survive. In addition to straightforward online accessibility over the long term, databases should ideally be regularly updated to include the latest useful information. In order to make this process efficient and compatible with the reproducibility of the research projects that need the databases, these updates should be clearly versioned and previous releases archived for the long term. In addition, unique identifiers should be assigned to individual database entries and maintained persistently across all versions.

Second, the database should be easily searchable and retrievable. Most of those mentioned in this chapter can be accessed via a Graphical User Interface (GUI) developed to browse and search data easily, for instance by typing keywords in a search box, providing a query molecule in SMILES format or as a file, or by drawing compounds or molecular fragments within a molecular sketcher. Such interfaces are particularly efficient to search for information about a few given molecules and to display them in a well-designed graphical representation. However, such interfaces become inefficient when a project requires a large amount of data, which will eventually have to be analyzed by the user through dedicated scripts and programs. In these cases, the information should be searchable and massively retrievable by command lines, for example, with an API through specific search and download commands. Ideally, the whole database content should be downloadable for local use by classic database management systems, such as MySQL or PostgreSQL, in order to be easily deployed and managed on the computers of advanced users.

Third, CADD databases and datasets should use renowned and well-accepted formats to store and deliver molecules to the users. As mentioned above, several strings and file formats are already available for this purpose, including SMILES, InChI, SDF, Mol, Mol2, PDB, and mmCIF. These formats are readily processed by most CADD software, making the use of the databases or datasets content straightforward.

Fourth, to make the interoperability between databases easier, they should include as much as possible well-accepted unique identifiers from long-standing key players in the field. For instance, the UniProt [81] ID provides a valuable solution to identify proteins. In addition, small molecules can be identified in many cases by one of the identifiers present in UniChem. This does not prevent the authors of new databases to create their own unique identifiers, for more flexibility. For example, ChEMBL uses its own unique identifier for proteins and ensures interoperability with other resources by providing a file mapping these ChEMBL IDs with UniProt [81] IDs.

Fifth, accurate information regarding the origin of the data stored in the database or dataset should be provided, as well as a detailed description of the manual or automatic curation processes applied to it.

Sixth, databases and datasets should have a clear usage license. Free- and open-access resources are often favored in academic environment, where funding may be limited, because they increase the visibility, maximize the use and impact of data, and facilitate the reuse of research results (Table 1.1).

Table 1.1 List of databases and datasets, along with their main usage and URL. When appropriate, the key purpose is reminded: training and validation of new approaches, or applicative usage. VS: virtual screening.

Name	Main usages	Description	Availability/URL	References
Databases of experimentally determined 3D structures of biomacromolecules and related resources				
PDBe	Docking Structure-based VS Target prediction Binding free energy estimation (Application, training, and validation)	As a member of the wwPDB, PDBe collects, organizes, and disseminates data on biological macromolecular structures. Contains more than 190,000 entries.	Can be freely searched here: https://www.ebi.ac.uk/pdbe REST API: https://www.ebi.ac.uk/pdbe/pdbe-rest-api Can be downloaded here: https://www.ebi.ac.uk/pdbe/services/ftp-access	[42]
PDB-Redo	Docking Structure-based VS Target prediction Binding free energy estimation (Application, training, and validation)	The PDB–REDO databank contains optimized versions of existing PDB entries with electron density maps, a description of model changes, and a wealth of model validation data.	Can be freely searched here: https://pdb-redo.eu API and download here: https://pdb-redo.eu/download-info.html	[71, 72]
Chemical Component Dictionary	Docking Ligand-based VS Structure-based VS (Application, training, and validation)	External reference file describing all residue and small molecule components found in PDB entries, maintained by the wwPDB Foundation.	Freely accessible here: https://www.wwpdb.org/data/ccd	[82]
Ligand Expo	Docking Ligand-based VS Structure-based VS (Application, training, and validation)	Provides chemical and structural information about small molecules within the structure entries of the Protein Data Bank (about 37,000 as of 2022). Maintained by the RCSB.	Freely accessible here: http://ligand-expo.rcsb.org Downloadable here in mmCIF, SDF, MOL, PDB, SMILES, and InChi: http://ligand-expo.rcsb.org/ld-download.html	[83]

(*continued*)

Table 1.1 (Continued)

Name	Main usages	Description	Availability/URL	References
PDBeChem	Docking Ligand-based VS Structure-based VS (Application, training, and validation)	Provides chemical and structural information about small molecules within the structure entries of the Protein Data Bank (more than 38,000 as of 2022). Maintained by PDB Europe.	Freely accessible here: https://www.ebi.ac.uk/pdbe-srv/pdbechem/	[84]
Databases of modeled 3D structures of biomacromolecules				
AlphaFold Protein Structure Database	Docking Structure-based VS (Application)	AlphaFold DB provides 200 million protein 3D structures predicted by AlphaFold, covering the proteomes of 48 organisms including humans.	Can be freely searched here: https://alphafold.ebi.ac.uk Sets of models can be downloaded here: https://alphafold.ebi.ac.uk/download	[75, 76]
ModBase	Docking Structure-based VS (Application)	Database of annotated comparative protein structure models obtained using the MODELLER program.	Can be freely searched here: https://modbase.compbio.ucsf.edu	[74]
SWISS-MODEL Repository	Docking Structure-based VS (Application)	Database of annotated 3D protein structure models generated by the SWISS-MODEL homology-modeling pipeline. Contains 2,250,005 models from SWISS-MODEL for UniProtKB targets as well as 180,763 structures from PDB with mapping to UniProtKB.	Can be freely searched here: https://swissmodel.expasy.org/repository	[73]

Databases of experimentally determined 3D structures of small molecules				
Cambridge Structure Database (CSD)	Ligand-based VS Structure-based VS	The CSD repository contains over one million accurate 3D small molecules of organic and metal–organic structures from x-ray and neutron diffraction analysis. Simple search is free, more advanced options require a license.	Freely accessible here: https://www.ccdc.cam.ac.uk/solutions/csd-core/components/csd/	[85]
COD	Ligand-based VS Structure-based VS	COD (Crystallography Open Database) provides a collection of 491,107 crystal structures of organic, inorganic, metal–organic compounds, and minerals, excluding biopolymers.	Freely accessible here: http://www.crystallography.net/cod	[86]
Data and information on proteins				
UniProtKB/Swiss-Prot	Target prediction Target validation	UniProtKB/Swiss-Prot is a manually annotated, nonredundant protein sequence database to provide all known relevant information about a particular protein. By combining numerous resources, the database became one of the major tools for biomedical research and drug target identification.	Can be freely searched here: https://www.uniprot.org Can be downloaded freely here: https://www.uniprot.org/uniprotkb?query=*	[81]

(*continued*)

Table 1.1 (Continued)

Name	Main usages	Description	Availability/URL	References
neXtProt	Target prediction Target validation	neXtProt is a comprehensive human-centric discovery platform, offering its users a seamless integration and navigation through protein-related data, for instance, function relationships with other diseases and molecular partners like drugs or chemicals. A section, in particular, is dedicated to protein–protein and protein–drug interaction data.	Can be freely searched here: https://www.nextprot.org	[87]
TCRD/Pharos	Ligand-based VS Structure-based VS Target prediction Binding free energy estimation (Application, training, and validation)	The Target Central Resource Database (TCRD) contains information about human targets, with special emphasis on poorly characterized proteins that can potentially be modulated using small molecules or biologics. Pharos is the web interface.	Freely accessible here: https://pharos.nih.gov/ TCRD can be downloaded here: http://juniper.health.unm.edu/tcrd/download/	[56]
Data and information on drugs				
CancerDrugs_DB	Licensed cancer drugs	Open access database of licensed cancer drugs with links to DrugBank and ChEMBL. IDs as well as information on targets and associated disease.	Freely accessible here: http://www.redo-project.org/cancer-drugs-db/ A machine-readable version of this database can be downloaded here: https://acfdata.coworks.be/cancerdrugsdb.txt The ReDO database of repurposing candidates in oncology can be accessed here: https://www.anticancerfund.org/en/redo-db	[88]

DrugCentral	Target prediction Drug repurposing	DrugCentral provides information on active ingredients’ chemical entities, pharmaceutical products, drug mode of action, indications, and pharmacologic action. Among others, sex-specific adverse effects are incorporated from FAERS database.	Can be freely searched here: https://drugcentral.org The database is available via Docker container: https://dockr.ly/35G46a6 and public instance drugcentral:unmtid-dbs.net:5433 A Python API is also available at: https://bit.ly/2RAHRtV.	[89]
Drug Repurposing Hub	Ligand-based VS Structure-based VS Target prediction Drug repurposing	Curated and annotated dataset of FDA-approved drugs, clinical candidates, and preclinical compounds with the accompanying information about their mechanism of action, protein targets as well as vendor’s ID. It currently stores information for 6807 compounds.	Freely accessible here: https://firedb.bioinfo.cnio.es/ The dataset can be downloaded at https://clue.io/repurposing#download-data	[90]
DrugBank	Ligand-based VS Structure-based VS Target prediction	DrugBank is a comprehensive database containing 2726 approved small molecule drugs, 1520 approved biologics (proteins, peptides, vaccines, and allergenic), 132 nutraceuticals, and over 6693 experimental (discovery-phase) drugs for a total of 14,665 drug entries. Additionally, 5278 nonredundant protein are linked to these drug entries.	Freely accessible here: https://go.drugbank.com	[52]

(*continued*)

Table 1.1 (Continued)

Name	Main usages	Description	Availability/URL	References
KEGG DRUG	Ligand-based VS Structure-based VS Target prediction	Comprehensive drug information resource for approved drugs in Japan, USA, and Europe unified based on the chemical structure and/or the chemical components of active ingredients. It contains 11,892 entries, including 5169 with human gene targets.	Freely accessible here: https://www.genome.jp/kegg/drug	[91]
TTD Therapeutics Target Database	Docking Structure-based VS Target prediction (Application, training, and validation)	A comprehensive collection of drugs with their corresponding targets. The database provides crosslinks to the target structure in PDB and Alphafold. Target sequences and structures are also available.	Accessible through login at: http://db.idrblab.net/ttd/	[92]
Databases of natural compounds				
COCONUT	Natural product database Virtual screening	COCONUT (COlleCtion of Open Natural ProdUcTs) online is an open-source project for Natural Products (NPs) storage, search, and analysis. It gathers data from over 50 open NP resources and is available free of charge and without any restriction. Each entry corresponds to a “flat” NP structure and is associated, when available, to their known stereochemical forms, literature, organisms that produce them, natural geographical presence, and diverse precomputed molecular properties.	https://coconut.naturalproducts.net	[93]

PSC-db	Natural product database Ligand-based	PSC-db, a unique plant metabolite database that categorizes the diverse phytochemical spaces by providing 3D-structural information along with physicochemical and pharmaceutical properties of the most relevant natural products.	http://pscdb.appsbio.utalca.cl	[94]
Super Natural II	Natural product database Ligand-based Toxicity	The database contains 325,508 natural compounds (NCs), including information about the corresponding 2D structures, physicochemical properties, predicted toxicity class, and potential vendors.	https://bioinf-applied.charite.de/supernatural_new/index.php	[95]
Databases of small molecules				
ChEBI	Ligand-based VS Structure-based VS	ChEBI (Chemical Entities of Biological Interest) is a freely available dictionary of about 122,000 molecular entities focused on "small" chemical compounds.	Freely browsable at https://www.ebi.ac.uk/chebi SDF files here: https://ftp.ebi.ac.uk/pub/databases/chebi/SDF and database files here: https://ftp.ebi.ac.uk/pub/databases/chebi	[96]
ChEMBL	Ligand-based VS Structure-based VS Target prediction Binding free energy estimation (Application, training, and validation)	Database containing 2.3 million small molecules and their experimentally measured activities on 14,000 protein targets and 2000 cells, extracted from 1.5 million assays.	https://www.ebi.ac.uk/chembl Freely accessible here: Downloadable in multiple formats: https://chembl.gitbook.io/chembl-interface-documentation/downloads	[48, 49]

(*continued*)

Table 1.1 (Continued)

Name	Main usages	Description	Availability/URL	References
ChemSpider	Ligand-based VS Structure-based VS	Collection of 115 million chemical structures compiled by the Royal Society of Chemistry from 277 data sources (e.g. DrugBank, BindingDB, ChEBI, vendors, etc.). It includes the conversion of chemical names to chemical structures, the generation of SMILES and InChI strings, as well as the prediction of many physicochemical parameters.	Freely searchable here: http://www.chemspider.com/Default.aspx	[69]
DrugSpaceX	Ligand-based VS Structure-based VS (Application)	101 million chemical products for virtual screening based on transformation rules with approved drug molecules as the starting points.	Freely accessible here: https://drugspacex.simm.ac.cn	[97]
FireDB	Docking Binding site prediction (Application, training, and validation)	Database of small molecule ligands and related binding residues part of a functional site. The database can be accessed by PDB codes or UniProt accession numbers.	Can be freely downloaded here: http://firedb.bioinfo.cnio.es/repository/current_FireDB_release_mysqldump/current_release.tgz	[98]
GDB-17	Ligand-based VS	GDB-17 enumerates 166.4 billion organic molecules up to 17 atoms of C, N, O, S, and halogens. Smaller sets of 50 million molecules or 11 million lead-like compounds are also available in SMILES format.	Freely accessible here: https://zenodo.org/record/7041051#.Y00Xcy0RqFo Smaller sets are available here: https://gdb.unibe.ch/downloads/	[53]
PubChem	Ligand-based VS Structure-based VS Target prediction (Application, training, and validation)	Open chemistry database at the NIH containing 112 million compounds and 301 million bioactivities, with information on chemical structures, identifiers, chemical and physical properties, biological activities, patents, health, safety, and toxicity data.	Freely accessible here: https://pubchem.ncbi.nlm.nih.gov Bulk downloads are possible from outputs or by FTP: https://ftp.ncbi.nlm.nih.gov/pubchem	[50]

SCUBIDOO	Ligand-based VS Structure-based VS (Application)	SCUBIDOO (Screenable Chemical Universe Based on Intuitive Data OrganizatiOn) 21 million virtual products originating from a small library of building blocks and a collection of organic reactions. The dataset is distributed in three representative and computationally tractable samples denoted as S, M, and L, containing 9994, 99,977, and 999,794 products, respectively.	Freely accessible here: https://scubidoo.pharmazie.uni-marburg.de/index.php Set download: https://scubidoo.pharmazie.uni-marburg.de/view/download.php	[99]
Zinc	Ligand-based VS Structure-based VS (Application)	Database of commercially available compounds for virtual screening. It contains 1.3 billion molecules, sourced from 310 catalogs from 150 vendors, with 2D and (for most) 3D structures. Of the 736 million lead-like molecules following the rule-or-four, 509 million are available for download in 3D ready for docking.	Freely accessible and downloadable here: https://zinc.docking.org https://zinc21.docking.org	[51]
Target-class centric database				
BiasDB	Target-class centric database	Manually curated database containing all published biased GPCR ligands.	Freely accessible here: https://biasdb.drug-design.de/	[100]
GLASS	Target-class centric database	GLASS (GPCR-Ligand Association) database is a manually curated repository for experimentally validated GPCR-ligand interactions. Contains 3056 GPCR (including 825 human ones) and 342,539 ligand entries.	Freely accessible here: https://zhanggroup.org/GLASS	[101]

(*continued*)

Table 1.1 (Continued)

Name	Main usages	Description	Availability/URL	References
GPCRdb	Target-class centric database	GPCRdb contains all human nonolfactory GPCRs (and >27,000 orthologs) in inactive, intermediate and active states, G-proteins, and arrestins. It includes over 2000 drug and in-trial agents and nearly 200,000 ligands with activity and availability data.	Freely accessible here: https://gpcrdb.org/	[102]
KinCoRe	Target-class centric database	Provides data for protein kinase sequences, structures, and phylogeny. It contains a list of FDA-approved PK inhibitors with known structures.	Can be freely searched here: http://dunbrack.fccc.edu/kincore Can be downloaded here: http://dunbrack.fccc.edu/kincore/download	[103]
KLIFS	Target-class centric database	KLIFS (Kinase–Ligand Interaction Fingerprints and Structures) contains over 5200 annotated kinase structures comprising 307 unique kinases and more than 3300 unique inhibitors, to support structure-based kinase research.	Freely accessible here: https://klifs.net	[104]
PDEStrIAn	Target-class centric database	PDEStrIAn (PhosphoDiEsterase Structure and ligand Interaction Annotated database) is a curated and annotated database of structures of catalytic PDE domains and inhibitors, collecting 377 PDB entries and 288 unique ligands.	Freely accessible here: http://pdestrian.vu-compmedchem.nl	[105]

Datasets for binding free energy estimation				
BioLiP	Docking Structure-based VS Binding site prediction	Semimanually curated database for high-quality, biologically relevant ligand–protein binding interactions. It contains 573,225 entries, involving 116,643 proteins from PDB and 327,620 ligands.	Can be freely searched here: https://zhanggroup.org/BioLiP/qsearch.html And downloaded here: https://zhanggroup.org/BioLiP/download.html	[106]
Binding MOAD	Binding free energy estimation (Training and validation)	High-quality ligand–protein structure database extracted from the PDB. Clearly identified biologically relevant ligands annotated with experimentally determined binding data extracted from literature. It contains 41,409 protein–ligand structures, 15,223 binding data, 20,387 different ligands, and 11,058 different families.	Freely accessible here: https://bindingmoad.org/ Different sets to download: https://bindingmoad.org/Home/download	[107, 108]
BindingDB	Binding free energy estimation (Training and validation)	Database of measured binding affinities, focusing chiefly on the interactions of proteins considered to be drug targets with drug-like small molecules. It contains 41,296 entries, involving 2,519,702 binding data for 8810 protein targets and 1,080,101 small molecules. BindingDB lists 5988 protein–ligand crystal structures with affinity measurements for proteins with 100% sequence identity, and 11,442 crystal structures allowing proteins to have 85% sequence identity.	Freely accessible here: https://www.bindingdb.org/rwd/bind/aboutus.jsp Can be freely downloaded, mainly in SDF format, at https://www.bindingdb.org/rwd/bind/chemsearch/marvin/SDFdownload.jsp?all_download=yes	[109]

(*continued*)

Table 1.1 (Continued)

Name	Main usages	Description	Availability/URL	References
PDBbind	Binding free energy estimation (Training and validation)	Comprehensive collection of experimentally measured binding affinity data for all biomolecular complexes deposited in the Protein Data Bank. It provides binding affinity data for a total of 23,496 biomolecular complexes, including protein–ligand (19,443), protein–protein (2852), protein–nucleic acid (1052), and nucleic acid–ligand complexes (149).	Can be freely searched here: http://www.pdbbind.org.cn/browse.php Can be freely downloaded here, after registration: http://www.pdbbind.org.cn/download.php	[110, 111]
Benchmark datasets				
CCD/Astex Validation Set	Docking (Validation/benchmarking)	Test set of 85 diverse, high-quality ligand–protein complexes from the PDB, for the validation of protein–ligand docking performance.	Freely accessible here: https://www.ccdc.cam.ac.uk/support-and-resources/Downloads/?d=27	[46]
CrossDocked2020	Structure-based VS (Training and validation)	22.6 million poses of 13,839 ligands (41.9% with affinity data) cross-docked into 2922 binding pockets across the Protein Data Bank.	Freely accessible here: https://github.com/gnina/models	[112]
D3R Grand Challenges	Binding free energy estimation Docking (Validation/benchmarking)	Collection of ligand–protein datasets used to benchmark docking software and binding free energy estimators, originally in a blind test. Collections are still available for a posteriori benchmarking.	Freely accessible here: https://drugdesigndata.org/about/grand-challenge	[113]

DEKOIS	Ligand-based VS Structure-based VS (Training and validation)	DEKOIS (Demanding Evaluation Kits for Objective In silico Screening) 2.0 library includes 81 high-quality benchmark sets for 80 protein targets. Positives were taken from BindingDB. Each positive is matched by 30 structurally diverse negatives with similar physicochemical properties.	Datasets free available per target available in SDF format at http://www.pharmchem.uni-tuebingen.de/dekois Full dataset here: http://www.pharmchem.uni-tuebingen.de/dekois/data/DEKOIS2.0_library/DEKOIS2.0_library.rar	[55]
DISCO	Structure-based VS (Training and validation)	Benchmark set for cross-docking using the targets listed in DUD-E. The completed benchmark contains 4399 ligand and receptor structures homologous to one of 95 targets, an average of 46 ligands per target.	Freely accessible here: http://disco.csb.pitt.edu/	[114]
DUD-E	Ligand-based VS Structure-based VS (Training and validation)	22,886 active compounds and their affinities against 102 targets + 50 decoys for each active having similar physicochemical properties but dissimilar 2D topology. Possibility to create decoys for user-defined ligands.	Freely accessible here: http://dude.docking.org All set archive download: http://dude.docking.org/db/subsets/all/all.tar.gz	[54]
Iridium	Docking (Validation/benchmarking)	Dataset of highly trustworthy protein–ligand 3D structures including a set of 121 structures named **Iridium-HT** for highly trustworthy and a second set of 104 structures named **Iridium-MT** for moderately trustworthy that violated some of the quality criteria. The datasets are freely available to download after registration.	Freely accessible here: https://www.eyesopen.com/iridium-database	[47]

(*continued*)

Table 1.1 (Continued)

Name	Main usages	Description	Availability/URL	References
LIT-PCBA	Ligand-based VS Structure-based VS (Training and validation)	PubChem Bioassay data-based set designed to incorporate actives and decoys with similar molecular properties. The dataset comprises 15 target collections with 9780 high-confidence actives and 407,839 unique inactives in total.	Freely accessible here: https://drugdesign.unistra.fr/LIT-PCBA/	[115]
Database of compounds IDs in some of the main small-molecule databases				
UniChem	Diverse	UniChem is large-scale nonredundant database of pointers between chemical structures and different databases and resources, including PubChem, ChEMBL, ZINC, BindingDB, or SwissLipids.	https://www.ebi.ac.uk/unichem/	[65]
Databases for ligand design				
sc-PDB-Frag	Ligand design	Database of protein-bound fragments for selecting bioisosteric scaffolds. It contains 12,000 fragments within 8077 ligand–protein complexes from the PDB, involving 2377 proteins and 5233 ligands.	Freely searchable at: http://bioinfo-pharma.u-strasbg.fr/scPDBFrag	[116]
SwissBioisostere	Ligand design	Open access database of >25 million unique molecular replacements with data on bioactivity, physicochemistry, chemical, and biological contexts extracted from the literature and related resources.	Freely searchable at: http://www.swissbioisostere.ch	[117]

Databases of binding sites				
M-CSA	Binding site prediction	CSA (Catalytic Site Atlas) lists enzyme active sites and catalytic residues in enzymes of 3D structure. It contains 1003 hand-curated entries, with detailed mechanistic descriptions. The entries in M-CSA represent 895 EC numbers, 73,211 SwissProt sequences, and 15,541 PDB files.	Can be freely searched here https://www.ebi.ac.uk/thornton-srv/m-csa/search And downloaded here https://www.ebi.ac.uk/thornton-srv/m-csa/download	[118]
PoSSuM	Docking Ligand design Binding site prediction	Database of 515,920 known and 9,160,203 putative ligand binding sites found in the Protein Data Bank (PDB).	Search mode for finding similar binding sites to a known ligand-binding site: https://possum.cbrc.jp/PoSSuM/search_k.html Search mode for predicting ligands that potentially bind to a structure of interest: https://possum.cbrc.jp/PoSSuM/search_p.html	[119, 120]
ProBiS-Dock Database	Docking Binding site prediction	Repository of 1,406,999 small-ligand binding sites.	Freely accessible here: http://probis-dock-database.insilab.org Freely accessible here: http://probis-dock-database.insilab.org/datasets	[121]

(*continued*)

Table 1.1 (Continued)

Name	Main usages	Description	Availability/URL	References
Datasets and databases related to ADME				
B3DB	ADME	Benchmark dataset for Blood-Brain Barrier permeability prediction, compiled from 50 published resources and containing numerical logBB values for 1058 compounds, and categorical BBB permeability labels (BBB+ or BBB-) for 7807 compounds.	Freely downloadable here: https://github.com/theochem/B3DB	[122]
HMDB	ADME	The Human Metabolome Database (HMDB) is a freely available electronic database containing detailed information about small molecule metabolites found in the human body. It is intended to be used for applications in metabolomics, clinical chemistry, biomarker discovery, and general education. The database is designed to contain or link three kinds of data: (i) chemical data, (ii) clinical data, and (iii) molecular biology/biochemistry data. The database contains 220,945 metabolite entries including both water-soluble and lipid-soluble metabolites. Additionally, 8610 protein sequences (enzymes and transporters) are linked to these metabolite entries.	Freely downloadable here: https://hmdb.ca Downloads in FASTA, SDF, XML format here: https://hmdb.ca/downloads	[123]

iCYP-MFE	ADME	Dataset of human Cytochrome P450 inhibitors for CYP1A2 (4471 inhibitors and 4886 non-inhibitors), CYP2C9 (3036, 6208), CYP2C19 (4392, 5479), CYP2D6 (1858, 8426), and CYP3A4 (4635, 7076).	Freely downloadable here: https://github.com/mldlproject/2021-iCYP-MFE	[124]
MetaCyc	ADME	MetaCyc is a curated database of experimentally elucidated metabolic pathways involved in both primary and secondary metabolism, as well as associated metabolites, reactions, enzymes, and genes. The goal of MetaCyc is to catalog the universe of metabolism by storing a representative sample of each pathway. MetaCyc currently contains 2937 pathways, 17,780 reactions, and 18,124 metabolites.	Freely downloadable here: https://metacyc.org	[125]
Metrabase	ADME	The **Met**abolism and **Tra**nsport Data**base** (**Metrabase**) provides structured data on interactions between proteins and compounds related to their metabolic fate and transport across biological membranes. The current version includes knowledge about 20 transporters and 13 CYPs, 3437 compounds, which represent 11,662 interaction records from 1209 literature references.	Freely searchable here: https://www-metrabase.ch.cam.ac.uk The whole MySQL and different flat files here: https://www-metrabase.ch.cam.ac.uk/metrabaseui/pageview/download/	[126]

(*continued*)

Table 1.1 (Continued)

Name	Main usages	Description	Availability/URL	References
NCATS-CYP	ADME	Dataset of 5094 compounds with experimentally determined antagonistic activity on different Cytochrome P450 (1742, 1984, and 2105 actives on CYP2D6, CYP2C9, and CYP3A4, respectively).	Freely downloadable here: https://pubchem.ncbi.nlm.nih.gov/bioassay/1645840 https://pubchem.ncbi.nlm.nih.gov/bioassay/1645842 https://pubchem.ncbi.nlm.nih.gov/bioassay/1645841	[127]
NCATS PAMPA1	ADME	Dataset of 2528 compounds including 295 molecules with ‘low or moderate parallel artificial membrane permeability assay (PAMPA) permeability at pH 7.4 (i.e. log P_{eff} < 2.0) and 1739 compound with ‘high PAMPA permeability’ (i.e. log P_{eff} > 2.5).	Freely downloadable here: https://pubchem.ncbi.nlm.nih.gov/bioassay/1508612	[128, 129]
NCATS-RLM	ADME	Dataset of 752 compounds unstable ($t_{1/2} \leq 30$ min) in a rat liver microsome stability profiling assay and 1774 stable ones ($t_{1/2} > 30$ min).	Freely downloadable here: https://pubchem.ncbi.nlm.nih.gov/bioassay/1508591	[130]
SMARTCyp dataset	ADME	Dataset for the construction of CYP450 site of metabolism (SOM) predict models It contains experimental SOM for different isoforms easily browsed through substructure search or downloadable as SDF files.	Freely searchable here: https://smartcyp.sund.ku.dk/mol_to_som?prediction=Search	[131]

Tox21-CYP	ADME	Dataset of 7683 compounds with experimentally determined antagonistic activity on different Cytochrome P450 (2372, 2914, 2447, 1523, and 1999 actives on CYP2C9, CYP2C19, CYP1A2, CYP3A4, and CYP2D6, respectively).	Freely downloadable here: https://pubchem.ncbi.nlm.nih.gov/bioassay/1671198 https://pubchem.ncbi.nlm.nih.gov/bioassay/1671197 https://pubchem.ncbi.nlm.nih.gov/bioassay/1671199 https://pubchem.ncbi.nlm.nih.gov/bioassay/1671201 https://pubchem.ncbi.nlm.nih.gov/bioassay/1671196	
Wang et al.	ADME	Dataset of 2358 molecules with categorical BBB permeability labels (BBB+ or BBB−).	Freely available as Supplementary Information here: https://chemistry-europe.onlinelibrary.wiley.com/doi/10.1002/cmdc.201800533	[132]
Datasets and databases related to toxicity				
Alves et al.	Toxicity	Dataset of 387 unique compounds, including 260 skin sensitizers and 127 non-sensitizers.	Freely available as Supplementary Information at https://ars.els-cdn.com/content/image/1-s2.0-S0041008X14004529-mmc2.xlsx	[133]
AMED Cardiotoxicity Database	Toxicity	Database of 9259 hERG inhibitors (IC50≤10 μM) and 279,718 inactive compounds (IC50>10 μM). Ligands of some other ion channels are also reported, including Nav1.5, Kv1.5, and Cav1.2.	Currently freely searchable at https://drugdesign.riken.jp/hERGdb/ Could be fully downloadable in the future.	[134]

(*continued*)

Table 1.1 (Continued)

Name	Main usages	Description	Availability/URL	References
CarPred	Toxicity	Experimental dataset of hERG assay results from 2130 chemicals, which were carried out under the same conditions.	Chemical structures of all compounds and their experimental hERG activities are available upon request to the authors	[135]
Cheng et al. 2011	Toxicity	Dataset of 1571 diverse chemicals including 1217 positives and 354 negatives on the *Tetrahymena pyriformis* toxicity test. The dataset contains the chemical names, CAS numbers, SMILES, and pIGC$_{50}$ values.	Freely available as Supplementary Information at https://ars.els-cdn.com/content/image/1-s2.0-S0045653510013500-mmc1.xls	[136]
Cheng et al. 2012	Toxicity	Dataset of 1604 unique compounds classified as "ready biodegradability" (RB) or "not ready biodegradability" (NRB) according to the biological oxygen demand test.	Freely available as Supplementary Information at https://ndownloader.figstatic.com/files/4180324	[137]
CTD (Comparative Toxicogenomics Database)	Toxicity	CTD 2021 contains 45 million toxicogenomic relationships for 16,394 chemicals, 51,344 genes, 5507 phenotypes, 7247 diseases, and 163,541 exposure events, from 601 comparative species.	Freely downloadable here: http://ctdbase.org/downloads	[138]
DGIdb (Drug-Gene Interaction Database)	Toxicity	DGIdb 4.0 (May 2021) contains 100,273 interactions between 39,095 molecules and 4847 genes, including 54,591 drug–gene interactions.	Freely accessible here: https://www.dgidb.org Downloads at: https://www.dgidb.org/downloads	[139]

DILIrank	Toxicity	The DILIrank dataset consists of 1036 FDA-approved drugs that are divided into four classes according to their potential for causing drug-induced liver injury (DILI): three groups (vMost-, vLess-, and vNo-DILI concern) with confirmed causal evidence, including 192,278 and 312 drugs, respectively, and one additional group (ambiguous-DILI-concern) with causality undetermined, including 254 drugs.	Freely available as a xlsx file here: https://www.fda.gov/science-research/liver-toxicity-knowledge-base-ltkb/drug-induced-liver-injury-rank-dilirank-dataset	[140]
ECOTOX	Toxicity	The ECOTOXicology Knowledgebase (ECOTOX) is a source for locating single chemical toxicity data for aquatic life, terrestrial plants, and wildlife. It provides single-chemical ecotoxicity data for over 12,540 chemicals on 13,741 with over 1.1 million test results from over 53,000 references.	https://cfpub.epa.gov/ecotox	[141]
Fan et al.	Toxicity	Dataset of 641 diverse chemicals labeled as negative or positive according to the *in vivo* micronucleus assay results, i.e. compounds able or not to induce chromosomal damage or disrupt the cell division.	Freely available as Supplementary Information at https://www.rsc.org/suppdata/c7/tx/c7tx00259a/c7tx00259a2.xlsx	[142]
FDAMDD	Toxicity	Maximum recommended daily dose (MRDD) for 1216 pharmaceuticals.	Freely available in PubChem as provided by EPA DSSTox https://pubchem.ncbi.nlm.nih.gov/bioassay/1195	

(*continued*)

Table 1.1 (Continued)

Name	Main usages	Description	Availability/URL	References
hERGCentral	Toxicity	hERG inhibition data obtained from a primary screen against more than 300,000 structurally diverse compounds at 1 and 10 μM.	Freely downloadable at https://www.cambridgemedchemconsulting.com/news/index_files/81f15972727e1fe70ae7f37514bdab58-362.html or at https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/7BVDG8	[143]
Mazzatorta et al.	Toxicity	Dataset of 445 compounds with Lowest Observed Adverse Effect (LOAEL) values for oral rat chronic toxicity.	Freely available as Supplementary Information at https://pubs.acs.org/doi/suppl/10.1021/ci8001974/suppl_file/ci8001974_si_001.xls	[144]
T3DB	Toxicity	The Toxin and Toxin Target Database (T3DB), a.k.a. the Toxic Exposome Database, currently houses 3678 toxins, including pollutants, pesticides, drugs, and food toxins, which are linked to 2073 corresponding toxin target records. Altogether there are 42,374 toxin-target associations. Available as CSV files including SMILES, InChi, and SDF formats.	Freely downloadable here: http://www.t3db.ca/downloads	[145]

Tox21 challenge dataset	Toxicity		A library of several thousands of compounds, including environmental chemicals and drugs, screened against a panel of nuclear receptor (NR) and stress response (SR) pathway assays. NR data cover Aryl hydrocarbon receptor (950 positive and 7219 negative datapoints), aromatase (360, 6866), androgen receptor full length (380, 8982), androgen receptor LBD (303, 8296), estrogen receptor alpha full length (937, 6760), estrogen receptor alpha LBD and PPARγ (446, 8307). SR data cover nuclear factor (erythroid-derived 2)-like 2/antioxidant responsive element (1098, 6069), ATAD5 (338, 8753), heat shock factor response element (428, 7722), and mitochondrial membrane potential (1142, 7722), p53 (537, 8097). Data available in SMILES and SDF formats.	Freely downloadable here: https://tripod.nih.gov/tox21/challenge/data.jsp	[146]
Xu. et al.	Toxicity		Dataset containing 7617 diverse compounds, including 4252 mutagens and 3365 nonmutagens based on the Ames test.	Freely available as Supplementary Information at https://pubs.acs.org/doi/suppl/10.1021/ci300400a/suppl_file/ci300400a_si_001.xls	[147]

(*continued*)

Table 1.1 (Continued)

Name	Main usages	Description	Availability/URL	References
Zhu et al.	Toxicity	Dataset of 7385 compounds with their lethal dose (LD_{50}) in rat acute toxicity by oral exposure.	Chemical structures of all compounds and their experimental LD_{50} values are available upon request to the authors	[148]
Datasets of aggregators				
Aggregator Advisor	Aggregation prediction	Dataset of about 12,600 experimentally known aggregators from published sources.	Data are freely available in SMILES format at: http://advisor.docking.org/rawdata/aggpage.txt	[149]
ChemAgg	Aggregation prediction	Positive set of 12,119 known aggregators from Aggregator Advisor; negative set of 24,172 approved, experimental and investigational drugs taken from DrugBank and considered as non-aggregators.	Data freely available as a xlsx file, in Supplementary Information of the publication: https://pubs.acs.org/doi/suppl/10.1021/acs.jcim.9b00541/suppl_file/ci9b00541_si_002.xlsx	[150]
Other databases and datasets				
Google Patents	Patent	Gather and give access to more than 87 million patents and patent applications from 17 patent offices. It includes advanced search capability and translation.	https://patents.google.com	

LINCS	Mechanism of action/ side effects	The Library of Integrated Network-Based Cellular Signatures collects information about responses of cell lines to compound treatment. It currently stores information for 21,231 small molecule perturbagens.	LINCS Data Portal (small molecules): http://lincsportal.ccs.miami.edu/SmallMolecules/ LINCS Signature API: http://lincsportal.ccs.miami.edu/sigc-api/swagger-ui.html#/	[151]
PharmGKB	Target prediction Target validation	PharmGKB is a comprehensive resource that curates knowledge about the impact of genetic variation on drug response for clinicians and researchers. The current version includes knowledge about 746 drugs in 201 pathways involving 25,561 variants.	Freely accessible here: https://www.pharmgkb.org Different sets are downloadable: https://www.pharmgkb.org/downloads	[152]
SMPDB	Target prediction Target validation	SMPDB (The Small Molecule Pathway Database) is an interactive, visual database containing more than 30,000 small molecule pathways found in humans only. The majority of these pathways are not found in any other pathway database. SMPDB is designed specifically to support pathway elucidation and pathway discovery. For drugs in particular, both pharmacodynamic and pharmacokinetic pathways are described.	Freely accessible here: https://www.smpdb.ca	[153]
STITCH (Search Tool for Interacting Chemicals)	Understanding drug's cellular impact	Stitch 5.0 contains 367,000 protein–chemical interactions, covering 430,000 chemicals and 9.6 million proteins from 2031 organisms.	Freely accessible here: http://stitch.embl.de Networks and flat files are downloadable at: http://stitch.embl.de/cgi/download.pl	[154]

References

1 Yu, W. and MacKerell, A.D. (2017). Computer-aided drug design methods. *Methods in Molecular Biology* 1520: 85–106.

2 Talevi, A. (2018). Computer-aided drug design: an overview. *Methods in Molecular Biology* 1762: 1–19.

3 Frye, L., Bhat, S., Akinsanya, K., and Abel, R. (2021). From computer-aided drug discovery to computer-driven drug discovery. *Drug Discovery Today: Technologies* 39: 111–117.

4 Tautermann, C.S. (2020). Current and future challenges in modern drug discovery. *Methods in Molecular Biology* 2114: 1–17.

5 Shaker, B., Ahmad, S., Lee, J. et al. (2021). In silico methods and tools for drug discovery. *Computers in Biology and Medicine* 137: 104851.

6 Liu, X., Ijzerman, A.P., and van Westen, G.J.P. (2021). Computational approaches for de novo drug design: past, present, and future. *Methods in Molecular Biology* 2190: 139–165.

7 Agoni, C., Olotu, F.A., Ramharack, P., and Soliman, M.E. (2020). Druggability and drug-likeness concepts in drug design: are biomodelling and predictive tools having their say? *Journal of Molecular Modeling* 26: 120–111.

8 Gemma, S. (2020). Structure-based design of biologically active compounds. *Molecules* 25: 3115.

9 Scotti, L. and Scotti, M.T. (2020). Recent advancement in computer-aided drug design. *Current Pharmaceutical Design* 26: 1635–1636.

10 Chen, Y. and Kirchmair, J. (2020). Cheminformatics in natural product-based drug discovery. *Molecular Informatics* 39: e2000171.

11 Wang, A. and Durrant, J.D. (2022). Open-source browser-based tools for structure-based computer-aided drug discovery. *Molecules* 27: 4623.

12 Mouchlis, V.D. et al. (2021). Advances in de novo drug design: from conventional to machine learning methods. *International Journal of Molecular Sciences* 22: 1676.

13 Velmurugan, D., Pachaiappan, R., and Ramakrishnan, C. (2020). Recent trends in drug design and discovery. *Current Topics in Medicinal Chemistry* 20: 1761–1770.

14 Zagotto, G. and Bortoli, M. (2021). Drug design: where we are and future prospects. *Molecules* 26: 7061.

15 Doytchinova, I. (2022). Drug design-past, present, future. *Molecules* 27: 1496.

16 Kar, S. and Leszczynski, J. (2020). Open access in silico tools to predict the ADMET profiling of drug candidates. *Expert Opinion on Drug Discovery* 15: 1473–1487.

17 Kar, S., Roy, K., and Leszczynski, J. (2022). In silico tools and software to predict ADMET of new drug candidates. *Methods in Molecular Biology* 2425: 85–115.

18 Kirchmair, J. et al. (2015). Predicting drug metabolism: experiment and/or computation? *Nature Reviews. Drug Discovery* 14: 387–404.

19 van de Waterbeemd, H. and Gifford, E. (2003). ADMET in silico modelling: towards prediction paradise? *Nature Reviews. Drug Discovery* 2: 192–204.

20 Wilson, G.L. and Lill, M.A. (2011). Integrating structure-based and ligand-based approaches for computational drug design. *Future Medicinal Chemistry* 3: 735–750. https://doi.org/10.4155/fmc.11.18.

21 Śledź, P. and Caflisch, A. (2018). Protein structure-based drug design: from docking to molecular dynamics. *Current Opinion in Structural Biology* 48: 93–102.

22 Wang, X., Song, K., Li, L., and Chen, L. (2018). Structure-based drug design strategies and challenges. *Current Topics in Medicinal Chemistry* 18: 998–1006.

23 Maia, E.H.B., Assis, L.C., de Oliveira, T.A. et al. (2020). Structure-based virtual screening: from classical to artificial intelligence. *Frontiers in Chemistry* 8: 343.

24 Lima, A.N. et al. (2016). Use of machine learning approaches for novel drug discovery. *Expert Opinion on Drug Discovery* 11: 225–239.

25 Anighoro, A. (2022). Deep learning in structure-based drug design. *Methods in Molecular Biology* 2390: 261–271.

26 Kimber, T.B., Chen, Y., and Volkamer, A. (2021). Deep learning in virtual screening: recent applications and developments. *International Journal of Molecular Sciences* 22: 4435.

27 Jia, L. and Gao, H. (2022). Machine learning for in sSilico ADMET prediction. *Methods in Molecular Biology* 2390: 447–460.

28 Nag, S. et al. (2022). Deep learning tools for advancing drug discovery and development. *3 Biotech* 12: 110–121.

29 Rodríguez-Pérez, R., Miljković, F., and Bajorath, J. (2022). Machine learning in chemoinformatics and medicinal chemistry. *Annual Review of Biomedical Data Science* 5: 43–65.

30 Palazzesi, F. and Pozzan, A. (2022). Deep learning applied to ligand-based de novo drug design. *Methods in Molecular Biology* 2390: 273–299.

31 Xu, Y. (2022). Deep neural networks for QSAR. *Methods in Molecular Biology* 2390: 233–260.

32 Vamathevan, J. et al. (2019). Applications of machine learning in drug discovery and development. *Nature Reviews. Drug Discovery* 18: 463–477.

33 Hansch, C., Steward, A.R., and Iwasa, J. (1965). The correlation of localization rates of benzeneboronic acids in brain and tumor tissue with substituent constants. *Molecular Pharmacology* 1: 87–92.

34 Muratov, E.N. et al. (2020). QSAR without borders. *Chemical Society Reviews* 49: 3525–3564.

35 Zhao, L., Ciallella, H.L., Aleksunes, L.M., and Zhu, H. (2020). Advancing computer-aided drug discovery (CADD) by big data and data-driven machine learning modeling. *Drug Discovery Today* 25: 1624–1638.

36 Zhu, H. (2020). Big data and artificial intelligence modeling for drug discovery. *Annual Review of Pharmacology and Toxicology* 60: 573–589.

37 Brown, N. et al. (2020). Artificial intelligence in chemistry and drug design. *Journal of Computer-Aided Molecular Design* 34: 709–715.

38 Vogt, M. (2018). Progress with modeling activity landscapes in drug discovery. *Expert Opinion on Drug Discovery* 13: 605–615.

39 PDB consortium (1971). Crystallography: Protein Data Bank. *Nature: New Biology* 233: 223–223.

40 wwPDB consortium (2019). Protein Data Bank: the single global archive for 3D macromolecular structure data. *Nucleic Acids Research* 47: D520–D528.

41 Burley, S.K. et al. (2021). RCSB Protein Data Bank: powerful new tools for exploring 3D structures of biological macromolecules for basic and applied research and education in fundamental biology, biomedicine, biotechnology, bioengineering and energy sciences. *Nucleic Acids Research* 49: D437–D451.

42 Armstrong, D.R. et al. (2020). PDBe: improved findability of macromolecular structure data in the PDB. *Nucleic Acids Research* 48: D335–D343.

43 Bekker, G.-J. et al. (2022). Protein Data Bank Japan: celebrating our 20th anniversary during a global pandemic as the Asian hub of three dimensional macromolecular structural data. *Protein Science* 31: 173–186.

44 Lawson, C.L. et al. (2016). “EMDataBank unified data resource for 3DEM.” *Nucleic Acids Res.* 44: D396–D403. doi:10.1093/nar/gkv1126

45 Romero, P.R. et al. (2020). BioMagResBank (BMRB) as a resource for structural biology. *Methods in Molecular Biology* 2112: 187–218.

46 Hartshorn, M.J. et al. (2007). Diverse, high-quality test set for the validation of protein-ligand docking performance. *Journal of Medicinal Chemistry* 50: 726–741.

47 Warren, G.L., Do, T.D., Kelley, B.P. et al. (2012). Essential considerations for using protein–ligand structures in drug discovery. *Drug Discovery Today* 17: 1270–1281.

48 Mendez, D. et al. (2019). ChEMBL: towards direct deposition of bioassay data. *Nucleic Acids Research* 47: D930–D940.

49 Gaulton, A. et al. (2017). The ChEMBL database in 2017. *Nucleic Acids Research* 45: D945–D954.

50 Kim, S. et al. (2021). PubChem in 2021: new data content and improved web interfaces. *Nucleic Acids Research* 49: D1388–D1395.

51 Irwin, J.J. et al. (2020). ZINC20-A free ultralarge-scale chemical database for ligand discovery. *Journal of Chemical Information and Modeling* 60: 6065–6073.

52 Wishart, D.S. et al. (2018). DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Research* 46: D1074–D1082.

53 Ruddigkeit, L., van Deursen, R., Blum, L.C., and Reymond, J.-L. (2012). Enumeration of 166 billion organic small molecules in the chemical universe database GDB-17. *Journal of Chemical Information and Modeling* 52: 2864–2875.

54 Mysinger, M.M., Carchia, M., Irwin, J.J., and Shoichet, B.K. (2012). Directory of useful decoys, enhanced (DUD-E): better ligands and decoys for better benchmarking. *Journal of Medicinal Chemistry* 55: 6582–6594.

55 Bauer, M.R., Ibrahim, T.M., Vogel, S.M., and Boeckler, F.M. (2013). Evaluation and optimization of virtual screening workflows with DEKOIS 2.0—a public library of challenging docking benchmark sets. *Journal of Chemical Information and Modeling* 53: 1447–1462.

56 Sheils, T.K. et al. (2021). TCRD and Pharos 2021: mining the human proteome for disease biology. *Nucleic Acids Research* 49: D1334–D1346.

57 Keiser, M.J. et al. (2009). Predicting new molecular targets for known drugs. *Nature* 462: 175–181.

58 Lounkine, E. et al. (2012). Large-scale prediction and testing of drug activity on side-effect targets. *Nature* 486: 361–367.

59 Gfeller, D. et al. (2014). SwissTargetPrediction: a web server for target prediction of bioactive small molecules. *Nucleic Acids Research* 42: W32–W38.

60 Daina, A., Michielin, O., and Zoete, V. (2019). SwissTargetPrediction: updated data and new features for efficient prediction of protein targets of small molecules. *Nucleic Acids Research* 47: W357–W364.

61 Weininger, D. (1988). SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *Journal of Chemical Information and Computer Sciences* 28 (1): 31–36. https://doi.org/10.1021/ci00057a005.

62 Weininger, D., Weininger, A. & Weininger, J. L. SMILES. 2. Algorithm for generation of unique SMILES notation. (2002). https://doi.org/10.1021/ci00062a008

63 Heller, S.R., McNaught, A., Pletnev, I. et al. (2015). InChI, the IUPAC international chemical identifier. *Journal of Cheminformatics* 7: 23–34.

64 Heller, S., McNaught, A., Stein, S. et al. (2013). InChI - the worldwide chemical structure identifier standard. *Journal of Cheminformatics* 5: 7–9.

65 Chambers, J. et al. (2014). UniChem: extension of InChI-based compound mapping to salt, connectivity and stereochemistry layers. *Journal of Cheminformatics* 6: 43–10.

66 Cereto-Massagué, A. et al. (2015). Molecular fingerprint similarity search in virtual screening. *Methods* 71: 58–63.

67 Seidel, T., Schuetz, D.A., Garon, A., and Langer, T. (2019). The Pharmacophore concept and its applications in computer-aided drug design. *Progress in the Chemistry of Organic Natural Products* 110: 99–141.

68 Giordano, D., Biancaniello, C., Argenio, M.A., and Facchiano, A. (2022). Drug design by pharmacophore and virtual screening approach. *Pharmaceuticals (Basel)* 15: 646.

69 Pence, H. E. & Williams, A. (2010). ChemSpider: an online chemical information resource. *J. Chem. Educ.* 87 (11): 1123–1124. https://doi.org/10.1021/ed100697w

70 Bourne, P.E. et al. (1997). Macromolecular crystallographic information file. *Methods in Enzymology* 277: 571–590.

71 Joosten, R.P., Joosten, K., Cohen, S.X. et al. (2011). Automatic rebuilding and optimization of crystallographic structures in the Protein Data Bank. *Bioinformatics* 27: 3392–3398.

72 Joosten, R.P., Long, F., Murshudov, G.N., and Perrakis, A. (2014). The PDB_REDO server for macromolecular structure model optimization. *IUCrJ* 1: 213–220.

73 Bienert, S. et al. (2017). The SWISS-MODEL Repository-new features and functionality. *Nucleic Acids Research* 45: D313–D319.

74 Pieper, U. et al. (2014). ModBase, a database of annotated comparative protein structure models and associated resources. *Nucleic Acids Research* 42: D336–D346.

75 Jumper, J. et al. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature* 596: 583–589.

76 Varadi, M. et al. (2022). AlphaFold protein structure database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Research* 50: D439–D444.

77 Wilkinson, M.D. et al. (2016). The FAIR guiding principles for scientific data management and stewardship. *Scientific data* 3: 160018–160019.

78 Attwood, T.K., Agit, B., and Ellis, L.B.M. (2015). Longevity of biological databases. *EMBnet.journal* 21: 803.

79 Finkelstein, J., Guarino, J., Huo, X. et al. (2022). Exploring determinants of longevity of biomedical databases. *Studies in Health Technology and Informatics* 290: 135–139.

80 Imker, H.J. (2018). 25 Years of molecular biology databases: a study of proliferation, impact, and maintenance. *Frontiers in Research Metrics and Analytics* 3: 18.

81 UniProt Consortium (2021). UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Research* 49: D480–D489.

82 Westbrook, J.D. et al. (2015). The chemical component dictionary: complete descriptions of constituent molecules in experimentally determined 3D macromolecules in the Protein Data Bank. *Bioinformatics* 31: 1274–1278.

83 Feng, Z. et al. (2004). Ligand depot: a data warehouse for ligands bound to macromolecules. *Bioinformatics* 20: 2153–2155.

84 Dimitropoulos, D., Ionides, J., and Henrick, K. (2006). Using PDBeChem to search the PDB ligand dictionary. In: *Current Protocols in Bioinformatics* (ed. A.D. Baxevanis, R. Page, G.A. Petsko, et al.) 14.3.1–14.3.3.

85 Groom, C.R., Bruno, I.J., Lightfoot, M.P., and Ward, S.C. (2016). The Cambridge structural database. *Acta Crystallographica. Section B: Structural Science, Crystal Engineering and Materials* 72: 171–179.

86 Vaitkus, A., Merkys, A., and Gražulis, S. (2021). Validation of the crystallography open database using the crystallographic information framework. *Journal of Applied Crystallography* 54: 661–672.

87 Zahn-Zabal, M. et al. (2020). The neXtProt knowledgebase in 2020: data, tools and usability improvements. *Nucleic Acids Research* 48: D328–D334.

88 Pantziarka, P., Capistrano, I.R., De Potter, A. et al. (2021). An open access database of licensed cancer drugs. *Frontiers in Pharmacology* 12: 627574.

89 Avram, S. et al. (2021). DrugCentral 2021 supports drug discovery and repositioning. *Nucleic Acids Research* 49: D1160–D1169.

90 Corsello, S.M. et al. (2017). The drug repurposing hub: a next-generation drug library and information resource. *Nature Medicine* 23: 405–408.

91 Kanehisa, M., Furumichi, M., Tanabe, M. et al. (2017). KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Research* 45: D353–D361.

92 Zhou, Y. et al. (2022). Therapeutic target database update 2022: facilitating drug discovery with enriched comparative data of targeted agents. *Nucleic Acids Research* 50: D1398–D1407.

93 Sorokina, M., Merseburger, P., Rajan, K. et al. (2021). COCONUT online: collection of open natural products database. *Journal of Cheminformatics* 13: 2–13.

94 Valdés-Jiménez, A. et al. (2021). PSC-db: a structured and searchable 3D-database for plant secondary compounds. *Molecules* 26: 1124.

95 Banerjee, P. et al. (2015). Super natural II--a database of natural products. *Nucleic Acids Research* 43: D935–D939.

96 Hastings, J. et al. (2016). ChEBI in 2016: improved services and an expanding collection of metabolites. *Nucleic Acids Research* 44: D1214–D1219.

97 Yang, T. et al. (2021). DrugSpaceX: a large screenable and synthetically tractable database extending drug space. *Nucleic Acids Research* 49: D1170–D1178.

98 Maietta, P. et al. (2014). FireDB: a compendium of biological and pharmacologically relevant ligands. *Nucleic Acids Research* 42: D267–D272.

99 Chevillard, F. and Kolb, P. (2015). SCUBIDOO: a large yet screenable and easily searchable database of computationally created chemical compounds optimized toward high likelihood of synthetic tractability. *Journal of Chemical Information and Modeling* 55: 1824–1835.

100 Bermudez, M., Nguyen, T.N., Omieczynski, C., and Wolber, G. (2019). Strategies for the discovery of biased GPCR ligands. *Drug Discovery Today* 24: 1031–1037.

101 Chan, W.K.B. et al. (2015). GLASS: a comprehensive database for experimentally validated GPCR-ligand associations. *Bioinformatics* 31: 3035–3042.

102 Kooistra, A.J. et al. (2021). GPCRdb in 2021: integrating GPCR sequence, structure and function. *Nucleic Acids Research* 49: D335–D343.

103 Modi, V. and Dunbrack, R.L. (2022). Kincore: a web resource for structural classification of protein kinases and their inhibitors. *Nucleic Acids Research* 50: D654–D664.

104 Kanev, G.K., de Graaf, C., Westerman, B.A. et al. (2021). KLIFS: an overhaul after the first 5 years of supporting kinase research. *Nucleic Acids Research* 49: D562–D569.

105 Jansen, C. et al. (2016). PDEStrIAn: a phosphodiesterase structure and ligand interaction annotated database as a tool for structure-based drug design. *Journal of Medicinal Chemistry* 59: 7029–7065.

106 Yang, J., Roy, A., and Zhang, Y. (2013). BioLiP: a semi-manually curated database for biologically relevant ligand-protein interactions. *Nucleic Acids Research* 41: D1096–D1103.

107 Smith, R.D. et al. (2019). Updates to binding MOAD (mother of all databases): polypharmacology tools and their utility in drug repurposing. *Journal of Molecular Biology* 431: 2423–2433.

108 Ahmed, A., Smith, R.D., Clark, J.J. et al. (2015). Recent improvements to Binding MOAD: a resource for protein-ligand binding affinities and structures. *Nucleic Acids Research* 43: D465–D469.

109 Liu, T., Lin, Y., Wen, X. et al. (2007). BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities. *Nucleic Acids Research* 35: D198–D201.

110 Liu, Z. et al. (2017). Forging the basis for developing protein-ligand interaction scoring functions. *Accounts of Chemical Research* 50: 302–309.

111 Liu, Z. et al. (2015). PDB-wide collection of binding data: current status of the PDBbind database. *Bioinformatics* 31: 405–412.

112 Francoeur, P.G. et al. (2020). Three-dimensional convolutional neural networks and a cross-docked data set for structure-based drug design. *Journal of Chemical Information and Modeling* 60: 4200–4215.

113 Parks, C.D. et al. (2020). D3R grand challenge 4: blind prediction of protein-ligand poses, affinity rankings, and relative binding free energies. *Journal of Computer-Aided Molecular Design* 34: 99–119.

114 Wierbowski, S.D., Wingert, B.M., Zheng, J., and Camacho, C.J. (2020). Cross-docking benchmark for automated pose and ranking prediction of ligand binding. *Protein Science* 29: 298–305.

115 Tran-Nguyen, V.-K., Jacquemard, C., and Rognan, D. (2020). LIT-PCBA: an unbiased data set for machine learning and virtual screening. *Journal of Chemical Information and Modeling* 60: 4263–4273.

116 Desaphy, J. and Rognan, D. (2014). sc-PDB-Frag: a database of protein-ligand interaction patterns for bioisosteric replacements. *Journal of Chemical Information and Modeling* 54: 1908–1918.

117 Cuozzo, A., Daina, A., Perez, M.A. et al. (2022). SwissBioisostere 2021: updated structural, bioactivity and physicochemical data delivered by a reshaped web interface. *Nucleic Acids Research* 50: D1382–D1390.

118 Ribeiro, A.J.M. et al. (2018). Mechanism and catalytic site Atlas (M-CSA): a database of enzyme reaction mechanisms and active sites. *Nucleic Acids Research* 46: D618–D623.

119 Ito, J.-I., Ikeda, K., Yamada, K. et al. (2015). PoSSuM v.2.0: data update and a new function for investigating ligand analogs and target proteins of small-molecule drugs. *Nucleic Acids Research* 43: D392–D398.

120 Tsuchiya, Y. and Tomii, K. (2020). Structural modeling and ligand-binding prediction for analysis of structure-unknown and function-unknown proteins using FORTE alignment and PoSSuM pocket search. *Methods in Molecular Biology* 2165: 1–11.

121 Konc, J., Lešnik, S., Škrlj, B., and Janezic, D. (2021). ProBiS-Dock database: a web server and interactive web repository of small ligand-protein binding sites for drug design. *Journal of Chemical Information and Modeling* 61: 4097–4107.

122 Meng, F., Xi, Y., Huang, J., and Ayers, P.W. (2021). A curated diverse molecular database of blood-brain barrier permeability with chemical descriptors. *Sci Data* 8: 289–211.

123 Wishart, D.S. et al. (2022). HMDB 5.0: the human metabolome database for 2022. *Nucleic Acids Research* 50: D622–D631.

124 Nguyen-Vo, T.-H. et al. (2021). https://doi.org/10.1021/acs.jcim.1c00628). iCYP-MFE: identifying human cytochrome P450 inhibitors using multitask learning and molecular fingerprint-embedded encoding. *Journal of Chemical Information and Modeling* 62 (21): 5059–5068.

125 Caspi, R. et al. (2020). The MetaCyc database of metabolic pathways and enzymes – a 2019 update. *Nucleic Acids Research* 48: D445–D453.

126 Mak, L. et al. (2015). Metrabase: a cheminformatics and bioinformatics database for small molecule transporter data analysis and (Q)SAR modeling. *Journal of Cheminformatics* 7: 1–12.

127 Gonzalez, E. et al. (2021). Development of robust quantitative structure-activity relationship models for CYP2C9, CYP2D6, and CYP3A4 catalysis and inhibition. *Drug Metabolism and Disposition* 49: 822–832.

128 Sun, H. et al. (2017). Highly predictive and interpretable models for PAMPA permeability. *Bioorganic & Medicinal Chemistry* 25: 1266–1276.

129 Siramshetty, V. et al. (2021). Validating ADME QSAR models using marketed drugs. *SLAS Discovery* 26: 1326–1336.

130 Siramshetty, V.B. et al. (2020). Retrospective assessment of rat liver microsomal stability at NCATS: data and QSAR models. *Scientific Reports* 10: 20713–20714.

131 Olsen, L., Montefiori, M., Tran, K.P., and Jørgensen, F.S. (2019). SMARTCyp 3.0: enhanced cytochrome P450 site-of-metabolism prediction server. *Bioinformatics* 35: 3174–3175.

132 Wang, Z. et al. (2018). In silico prediction of blood-brain barrier permeability of compounds by machine learning and resampling methods. *ChemMedChem* 13: 2189–2201.

133 Alves, V.M. et al. (2015). Predicting chemically-induced skin reactions. Part I: QSAR models of skin sensitization and their application to identify potentially hazardous compounds. *Toxicology and Applied Pharmacology* 284: 262–272.

134 Sato, T., Yuki, H., Ogura, K., and Honma, T. (2018). Construction of an integrated database for hERG blocking small molecules. *PLoS One* 13: e0199348.

135 Lee, H.-M. et al. (2019). Computational determination of hERG-related cardiotoxicity of drug candidates. *BMC Bioinformatics* 20: 250–273.

136 Cheng, F. et al. (2011). In silico prediction of Tetrahymena pyriformis toxicity for diverse industrial chemicals with substructure pattern recognition and machine learning methods. *Chemosphere* 82: 1636–1643.

137 Cheng, F. et al. (2012). In silico assessment of chemical biodegradability. *Journal of Chemical Information and Modeling* 52: 655–669.

138 Davis, A.P. et al. (2021). Comparative Toxicogenomics Database (CTD): update 2021. *Nucleic Acids Research* 49: D1138–D1143.

139 Freshour, S.L. et al. (2021). Integration of the Drug-Gene interaction database (DGIdb 4.0) with open crowdsource efforts. *Nucleic Acids Research* 49: D1144–D1151.

140 Chen, M. et al. (2016). DILIrank: the largest reference drug list ranked by the risk for developing drug-induced liver injury in humans. *Drug Discovery Today* 21: 648–653.

141 Olker, J.H. et al. (2022). The ECOTOXicology knowledgebase: a curated database of ecologically relevant toxicity tests to support environmental research and risk assessment. *Environmental Toxicology and Chemistry* 41: 1520–1539.

142 Fan, D. et al. (2018). In silico prediction of chemical genotoxicity using machine learning methods and structural alerts. *Toxicology Research* 7: 211–220.

143 Du, F. et al. (2011). hERGCentral: a large database to store, retrieve, and analyze compound-human Ether-à-go-go related gene channel interactions to facilitate cardiotoxicity assessment in drug development. *Assay and Drug Development Technologies* 9: 580–588.

144 Mazzatorta, P., Estevez, M.D., Coulet, M., and Schilter, B. (2008). Modeling oral rat chronic toxicity. *Journal of Chemical Information and Modeling* 48: 1949–1954.

145 Wishart, D. et al. (2015). T3DB: the toxic exposome database. *Nucleic Acids Research* 43: D928–D934.

146 Huang, R. et al. (2016). Tox21Challenge to build predictive models of nuclear receptor and stress response pathways as mediated by exposure to environmental chemicals and drugs. *Frontiers in Environmental Science* 3: 85.

147 Xu, C. et al. (2012). In silico prediction of chemical Ames mutagenicity. *Journal of Chemical Information and Modeling* 52: 2840–2847.

148 Zhu, H. et al. (2009). Quantitative structure-activity relationship modeling of rat acute toxicity by oral exposure. *Chemical Research in Toxicology* 22: 1913–1921.

149 Irwin, J.J. et al. (2015). An aggregation advisor for ligand discovery. *Journal of Medicinal Chemistry* 58: 7076–7087.

150 Yang, Z.-Y. et al. (2019). Structural analysis and identification of colloidal aggregators in drug discovery. *Journal of Chemical Information and Modeling* 59: 3714–3726.

151 Stathias, V. et al. (2020). LINCS data portal 2.0: next generation access point for perturbation-response signatures. *Nucleic Acids Research* 48: D431–D439.

152 Whirl-Carrillo, M. et al. (2021). An evidence-based framework for evaluating pharmacogenomics knowledge for personalized medicine. *Clinical Pharmacology and Therapeutics* 110: 563–572.

153 Jewison, T. et al. (2014). SMPDB 2.0: big improvements to the small molecule pathway database. *Nucleic Acids Research* 42: D478–D484.

154 Szklarczyk, D. et al. (2015). STITCH 5: augmenting protein–chemical interaction networks with tissue and affinity data. *Nucleic Acids Research* 44: D380–D384.