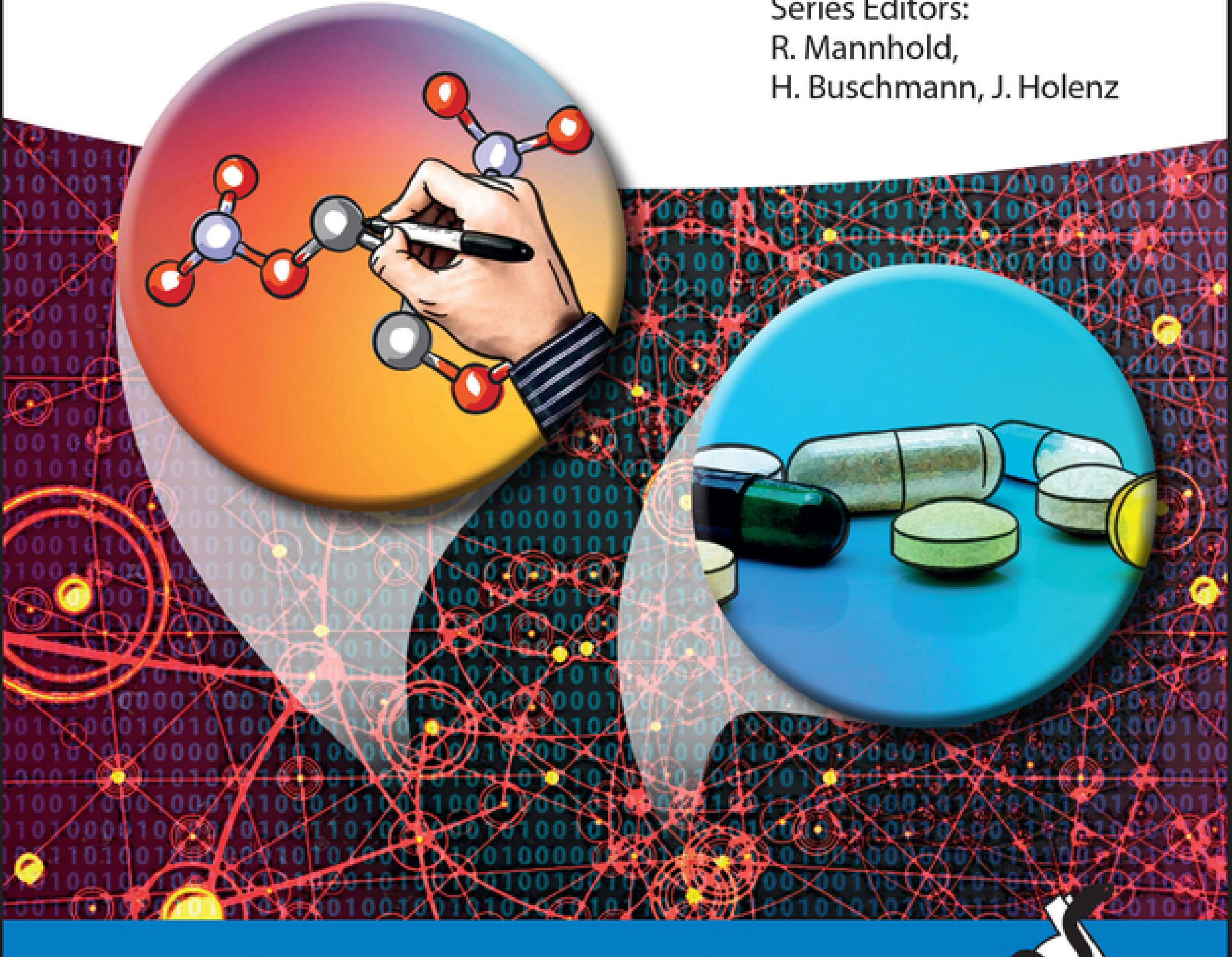
WILEY-VCH

Edited by Antoine Daina, Michael Przewosny, and Vincent Zoete

Open Access Databases and Datasets for Drug Discovery

Volume 83

Series Editors:
R. Mannhold,
H. Buschmann, J. Holenz



Methods and Principles in Medicinal Chemistry

Open Access Databases and Datasets for Drug Discovery

Methods and Principles in Medicinal Chemistry

Previous Volumes of the Series

Open Access Databases and Datasets for Drug Discovery

Edited by Antoine Daina, Michael Przewosny, and Vincent Zoete

WILEY VCH

Volume Editors

Antoine Daina
SIB Swiss Institute of Bioinformatics
1015 Lausanne
Switzerland

Michael Przewosny
Borngasse 43
52064 Aachen
Germany

Vincent Zoete
SIB Swiss Institute of Bioinformatics
UNIL University of Lausanne and
Ludwig Institute for Cancer Research
1015 Lausanne
Switzerland

Series Editors

Prof. Dr. Raimund Mannhold†
Rosenweg 7
40489 Düsseldorf
Germany

Dr. Helmut Buschmann
Sperberweg 15
52076 Aachen
Germany

Dr. Jörg Holenz
BIAL - Portela & C^{a}., S.A.
Av. Siderurgia Nacional
4745–457 Coronado
Portugal

Cover Design and Images: SCHULZ Grafik-Design

All books published by **WILEY-VCH** are carefully produced. Nevertheless, authors, editors, and publisher do not warrant the information contained in these books, including this book, to be free of errors. Readers are advised to keep in mind that statements, data, illustrations, procedural details or other items may inadvertently be inaccurate.

Library of Congress Card No.: applied for

British Library Cataloguing-in-Publication Data
A catalogue record for this book is available from the British Library.

Bibliographic information published by the Deutsche Nationalbibliothek
The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available on the Internet at <http://dnb.d-nb.de>.

Print ISBN: 978-3-527-34839-8
ePDF ISBN: 978-3-527-83047-3
ePub ISBN: 978-3-527-83048-0
oBook ISBN: 978-3-527-83049-7

Typesetting Straive, Chennai, India

Contents

Series Editors Preface

The work of natural scientists in all scientific disciplines has changed a lot in the recent decade. Access to information and data in scientific databases has become essential for effective and efficient work. In addition to the commercial databases from professional providers, open access databases from associations and institutes have also become increasingly popular for medicinal chemists in academia and pharmaceutical industry.

The latest volume of our book series entitled "Open Access Databases and Datasets for Drug Discovery" provides an exemplary overview of some of the most important databases and applications that should be of great help to the medicinal chemistry community as information source and motivation to explore the growing and existing field of open access databases and useful datasets. The book surely will support all type of scientists working in the field of drug discovery and medicinal chemistry who need information from databases to support their work.

It all started in the late 2010s when Raimund Mannhold suggested this topic in our annual editor meetings as a long-cherished heart's desire. And in 2019 he was successful to convince Antoine Daina and Vincent Zoete, who are well-known scientists in this field, to edit such a book.

After industrial practice as computational chemist for agrochemical research and academic experience as lecturer and researcher in drug discovery, Antoine joined the SIB Swiss Institute of Bioinformatics in 2012. He is now senior scientist in the Molecular Modeling Group in charge of methodological developments in the SwissDrugDesign program, of supporting drug discovery projects and of teaching computer-aided drug design.

Vincent joined the SIB Swiss Institute of Bioinformatics in 2004. He was the associate group leader of the SIB Molecular Modeling Group until 2017 and then group leader from 2017 until now. Besides this, Vincent is Associate Professor in molecular modeling at the University of Lausanne since 2022 and coordinator/developer of SwissDock.ch, SwissParam.ch, SwissBioisostere.ch, SwissTargetPrediction.ch, SwissSimilarity.ch, and SwissADME.ch.

The Swiss Institute of Bioinformatics hosting the Click2Drug Webpage provides the most comprehensive collection of worldwide available databases and application tools in the field of drug discovery.

At the same time Helmut Buschmann remembered his old colleague Michael Przewosny from our time together at Grünenthal GmbH located in Aachen. Michael has over 20 years of experience in pharmaceutical research and drug discovery. He held several positions as laboratory manager in medicinal chemistry and process development. Michael has created a competitive intelligence department at Grünenthal in Aachen, where he was responsible for such database and application tools as service for the entire research organization. It took some time to convince Michael for such a book assignment, but finally he was motivated to join the group of Antoine and Vincent.

Together they brought many years of experience in the development of such databases, reinforced by many years of experience in using such databases in the field of drug discovery.

Jointly Antoine, Vincent, and Michael started in late 2019 with a collection of ideas and agreed after long discussions on a useful structure of such a broad research area with an enormous rapid development.

After a successful start, there was now a long, rocky, and chaotic road ahead of them accompanied by the Covid pandemic. There were many disappointments, but they never gave up. They worked very hard and were always successful to find a way forward. Then another major setback followed. Raimund died unexpectedly after a short illness on October 14, 2022, and was not able to see the successful completion. We, the series editors and the publisher, are all the more pleased that the editors have dedicated this volume to his memory. Raimund accompanied the book series from the first volume published early as 1993 until his death and was able to enjoy the publishing of volume 81 in June 2022 "Flow and Microreactor Technology in Medicinal Chemistry" edited by Esther Alza, shortly before he passed away.

The editors managed to edit a book with the support of the best authors in the field to provide the interested reader with a detailed overview of open-access databases and datasets for drugs from early to late phases of the lengthy drug discovery process. In such rapidly growing research field, the picture of open databases and datasets remains always incomplete.

It is all the more important that the authors managed to edit a volume that depicts a wide variety of resources from the most generalist to most specialized ones. Such a volume can never be a complete and encyclopedic collection of all existing databases, but it acts much more like guidance and motivation to deal with such databases and the resulting possibilities. In different chapters the most relevant tools and databases and apps are described by explaining case studies and examples to get an easy and direct introduction to use these tools.

Antoine, Vincent, and Michael have managed with great passion to persuade and encourage the authors to provide as much practical advice as possible with step-by-step guides and helpful use cases for the interested reader of all disciplines involved in drug hunting, bringing new, powerful, and safe medicines to the patients. The selected and compiled data collection of databases and apps provides a strong comprehensive basis as a kind of guided tour through the very dense jungle of public available scientific information.

The editors have structured the guidance book in 3 thematic sections and 10 chapters. Antoine, Vincent, and Michael have not only edited the book but also contributed with their long experience and great knowledge as authors and co-authors of some of the chapters.

As a general introduction to the volume edited by Antoine and Vincent themselves with the support of their coworkers, a comprehensive overview to the topic and a rich annotated list of data sources entitled "Open Access Databases and Datasets for Computer-Aided Drug Design. A Short List Used in the Molecular Modelling Group of the SIB" is provided. The core of the book presented in part I and II consists of seven diverse and high-quality resources presented by their developers, categorized in small molecules or macromolecular targets and diseases.

Part I is dedicated to small molecules and contains three chapters describing the most popular databases in this field:

- PubChem: A Large-Scale Public Chemical Database for Drug Discovery, edited by Sunghwan Kim and Evan E. Bolton.
- DrugBank Online: A How-to Guide, edited by Christen M. Klinger, Jordan Cox, Denise So, Teira Stauth, Michael Wilson, Alex Wilson, and Craig Knox
- Bioisosteric Replacement for Drug Discovery Supported by the SwissBioisostere Database, edited by Antoine Daina, Alessandro Cuozzo, Marta A.S. Perez, and Vincent Zoete

Part II focuses on macromolecular targets and diseases comprising the following chapters:

- The Protein Data Bank (PDB) and Macromolecular Structure Data Supporting Computer-Aided Drug Design, edited by David Armstrong, John Berrisford, Preeti Choudhary, Lukas Pravda, James Tolchard, Mihaly Varadi, and Sameer Velankar
- The SWISS-MODEL Repository of 3D Protein Structures and Models, edited by Xavier Robin, Andrew Waterhouse, Stefan Bienert, Gabriel Studer, Leila T. Alexander, Gerardo Tauriello, Torsten Schwede, and Joana Pereira
- PDB-REDO in Computational-Aided Drug Design (CADD), edited by Ida de Vries, Anastassis Perrakis, and Robbie P. Joosten
- Pharos and TCRD: Informatics Tools for Illuminating Dark Targets, edited by Keith J. Kelleher, Timothy K. Sheils, Stephen L. Mathias, Dac-Trung Nguyen, Vishal Siramshetty, Ajay Pillai, Jeremy J. Yang, Cristian G. Bologa, Jeremy S. Edwards, Tudor I. Oprea, and Ewy Mathé

Part III of the book is dedicated to user's point of view working in academia and pharmaceutical industry with two chapters:

- Mining for Bioactive Molecules in Open Databases, edited by Guillem Macip, Júlia Mestres-Truyol, Pol Garcia-Segura, Bryan Saldivar-Espinoza, Santiago Garcia-Vallvé, and Gerard Pujadas
- Open Access Databases – An Industrial View, edited by Michael Przewosny

Overall, after a long and difficult journey an outstanding collection of database and dataset information is provided that will enable the interested reader an easy start to use such tools or to expand their scope by an extension of the previous application.

With this, we – the series editors – sincerely believe that readers would be highly benefited from the contents of this book.

We would like to thank Antoine, Vincent, and Michael to put the brilliant contributions of the authors together and to guide them through an adventurous journey; all authors for their brilliant contributions and their patience; and Frank Weinreich, Stefanie Volk, and their coworkers, especially Aswini M. from the content analysis and refinement team, for their great support to make this book finally possible.

Aachen, Porto, and Bonn, July 2023

Helmut Buschmann
Jörg Holenz
Christa Müller

Raimund Mannhold – A Personal Obituary from the Series Editors

Source: http://www.raimund-mannhold.de/curriculum-vitae/

Raimund Mannhold died on October 14, 2022, after a short and serious illness at the age of 74. Nevertheless, the news of his death came as a great surprise to his immediate family and to us. Raimund accompanied the book series "Methods and Principles in Medicinal Chemistry" from the first volume published as early as 1993 until his death and was able to enjoy the publishing of volume 81 in June 2022 entitled "Flow and Microreactor Technology in Medicinal Chemistry" edited by Esther Alza, shortly before he passed away.

Established in 1993, the series "Methods and Principles in Medicinal Chemistry" has become a crucial source of information within the medicinal chemistry community and beyond. Authors and editors of the series come from pharmaceutical industry as well as from academic institutions, fostering a more active exchange between these domains.

Over time, Raimund found support from a number of internationally renowned experts and entrepreneurs in medicinal chemistry. Povl Krogsgaard-Larsen, Hendrik Timmerman, Hugo Kubinyi, and Gerd Folkers as retired series editors had a decisive influence on the book series and, like Raimund, have contributed to it becoming a figurehead for medicinal chemistry worldwide.

The following picture shows Raimund (middle) with Gerd Folkers (left) and Hugo Kubinyi during the celebration of the 25th volume of the book series in 2005.

Source: Wiley-VCH

From the very beginning, the series focused on topical volumes covering hot concepts and technologies, and the reader will not miss any important topic in the field. The range of topics is as diverse as are the challenges facing modern drug developers, spanning the fields of organic chemistry, pharmacology, toxicology, life science, and analytics, the latter also including bioinformatics, chemoinformatics, and proteomics.

Raimund's heart beat for his book series, and he was now the only founding editor since the publication of the first volume 30 years ago (1993); now it must live on without Raimund, not as before, but it will continue to live on in order to preserve his legacy. That is the obligation of the current series editors Christa Müller, Jörg Holenz, and Helmut Buschmann.

Our common goal was to be able to celebrate volume 100 together; we were just able to publish volume 81 together, but without Raimund, without his commitment, and without his strong will to document the knowledge of medicinal chemistry of our time, it will be not an easy task to continue as usual. Without him there is a hard road ahead of us to fulfill his legacy and with great sadness to continue without him. But we see at the same time it as a great obligation to continue the book series in his spirit.

Raimund's life was shaped by pharmaceutical science. He was born in 1948 in Haltern (North Rhine Westphalia, Germany). From 1970 to 1973 he studied pharmacy at the Frankfurt University, received his doctorate in 1977 from the University of Düsseldorf, and in 1982 Raimund received the Venia Legendi for the subject

Physiology. In 1990 he was promoted to the professor of Molecular Drug Research at Heinrich-Heine University in Düsseldorf until his retirement on July 9, 2012.

The most important stages of his scientific career can be summarized as follows[1]:

1970–1973	Study of Pharmaceutical Sciences at the Johann-Wolfgang von Goethe Universität Frankfurt/Main
October 1973– September 1987	Scientific assistant at the Department of Clinical Physiology, Heinrich-Heine-Universität Düsseldorf
July 1977	PhD at the Department of Clinical Physiology (Heinrich-Heine-Universität Düsseldorf, Prof. Dr. R. Kaufmann). Thesis: Investigations on the Ca-antagonistic mode of action and the structure-activity relationships of verapamil
December 1982	Habilitation, conferred by the Medical Faculty of the Heinrich-Heine-Universität Düsseldorf. Title of monograph: Ca-antagonists of the aliphatic amine type and structurally related heart-active drugs – investigations on pharmacological and physicochemical properties
Since 1984	Contributing Editor of “Drugs of Today” and “Drugs of the Future”
January 1989– October 1990	Guest scientist at the Department for Pharmacochemistry, Vrije Universiteit, Amsterdam, NL (Prof. Dr. Henk Timmerman)
November 1990– July 2012	Professorship for Molecular Drug Research at the Heinrich-Heine-Universität Düsseldorf until his retirement on July 9, 2012
Since 1993	Editor of the book series “Methods and Principles in Medicinal Chemistry,” Wiley-VCH, Weinheim, together with Hugo Kubinyi and Henk Timmerman (and since 2001 with Gerd Folkers)
Since 2001	Regional editor of Mini-Reviews in Medicinal Chemistry
Since 2005	Editorial board member of Medicinal Chemistry and Current Computer-Aided Drug Design
October– November 2011	Visiting professor at the School of Pharmaceutical Sciences, University of Geneva, University of Lausanne, Switzerland

His work as the serial editor of his book series will perpetuate his memory in the pharmaceutical community worldwide. His book series has now established itself as an internationally recognized standard, and millions of scientists will continue to see his name and appreciate his works in the future.

With the death of Raimund we lose a part of the spirit of the book series, which is very difficult to get over. But his footprint on the volumes published so far will be documented forever and thus remain a valuable part of scholarship.

[1] http://www.raimund-mannhold.de/curriculum-vitae/

Dear Raimund, in addition to your content-related input, we will also miss the extremely precise planning for your beloved book series.

We promise to continue your and now our book series "Methods and Principles in Medicinal Chemistry" in your spirit, even beyond volume 100.

With deep sadness, but filled with the thought of carrying your spirit on,

Bonn, Porto, and Aachen, July 2023 *Christa, Jörg, and Helmut*

A Personal Foreword

When we think about computers to assist drug discovery, what comes to mind for most of us are the algorithms and graphics to calculate and visualize all sorts of molecular properties. What is less obvious is the knowledge that can be produced from the data itself. Today, a large amount and a vast diversity of data related to medicinal chemistry and drug discovery are available. With a few clicks, anyone can freely access downloadable raw datasets or browse more sophisticated structured databases.

This book aims to provide the reader with a detailed overview of open-access databases and datasets for drug discovery. While the picture is inevitably incomplete, it depicts a wide variety of resources from the most generalist to the most specialized. The volume begins with a rich annotated list of data sources considered of importance for (computer-aided) drug discovery and concludes with argued user perspectives. The core of the book consists of seven diverse and high-quality resources presented by their developers, categorized in *Small molecules* or *Macromolecular targets and diseases*. We have encouraged the authors to provide as much practical advice as possible with step-by-step guides and helpful use cases for medicinal chemists. Here we would like to express our deep gratitude to all the expert contributors for their remarkable commitment and admirable patience.

It all started in 2019 when the late Professor Raimund Mannhold contacted us for this project. The book is dedicated to his memory.

The process itself has been a long and chaotic journey through the COVID-19 pandemic.

Let us warmly thank the series editor, Dr. Helmut Buschmann and people at Wiley-VCH without whom this would not have been possible, in particular, Dr. Frank Weinreich, Stefany Volk, Satvinder Kaur, and Aswini Murugadass.

We wish you a pleasant and instructive reading!

Antoine Daina, Michael Przewosny, and Vincent Zoete

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*)