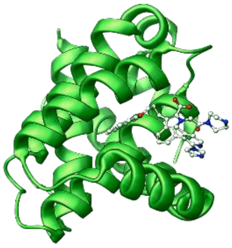
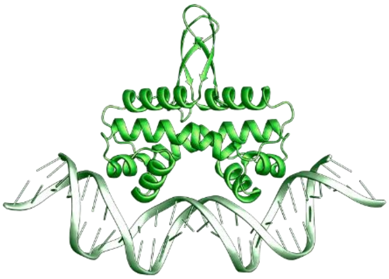
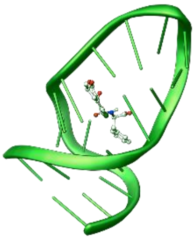
PART 1. Beginner's Guide to the PDBbind Database

The PDBbind database provides a comprehensive collection of experimental binding affinity data for the molecular complexes in the Protein Data Bank (PDB). This type of information is much needed by various computational and machine-learning studies on molecular interaction. The prototype of PDBbind was first released to the public in May 2004. Since 2007, this database has been updated basically on an annual base to keep up with the growth of PDB. **The latest release is version 2021.**


Protein-ligand complex

Protein-protein complex


Protein-nucleic acid complex


Nucleic acid-ligand complex

What does the PDBbind database provide?

- **Experimental binding data:** The main value of PDBbind is the collection of experimentally measured binding affinity data (K_d, K_i or IC_{50} values) that match the molecular complex structures in PDB. Molecular complexes under consideration include protein-ligand, protein-protein, protein-nucleic acid, and nucleic acid-ligand complexes. All binding data are curated by our team from peer-reviewed publications rather than being copied from third-party resources. A total of 45,200 publications have been checked for this purpose.

- **Processed complex structures:** As an important feature, PDBbind also provides carefully processed structure files for all protein-ligand complexes included in its contents, which can be readily utilized by most molecular modeling software. In brief, each complex structure is split into a protein molecule (in the PDB format) and a ligand molecule (in the Mol2 and SDF format). Atom/bond types on the ligand molecules are assigned properly by special computer programs and also confirmed by manual inspection. All processed protein-ligand structure files can be downloaded in a package.

- **Web-based functions:** The user can access PDBbind through a web portal called **PDBbind+**. On this web site, basic information of each complex is summarized on a single page. Text-based and structure-based search among the contents of PDBbind are also enabled. Besides, this web site also incorporates valuable functional modules that leverage the contents of PDBbind and the Protein Data Bank. On-line job computation are enabled with the support of more powerful cloud computing resources.

References and Notes

The PDBbind database is developed and maintained by Prof. Renxiao Wang's group at the School of Pharmacy, Fudan University in Shanghai, P. R. China. To cite the PDBbind database, please refer to the following publications:

(1) Liu, Z.H. et al. *Acc. Chem. Res.* 2017, *50*, 302-309. (PDBbind v.2016)
(2) Liu, Z.H. et al. *Bioinformatics*, 2015, *31*, 405-412. (PDBbind v.2014)
(3) Wang, R. X.; et al. *J. Med. Chem.* 2005, *48*, 4111-4119; *J. Med. Chem.* 2004, *47*, 2977-2980. (proto-type)

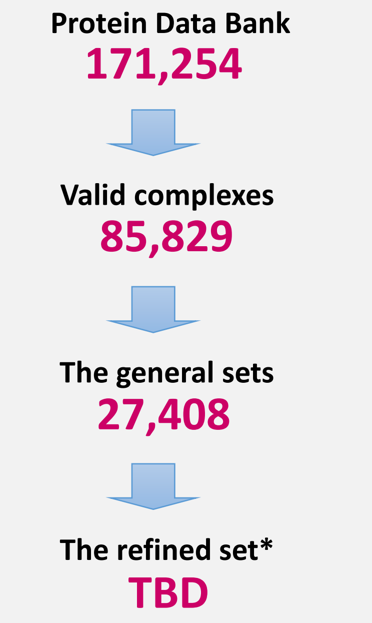
A brief history of the PDBbind database*

Version	Entries In PDB	All complex with binding data	Protein-ligand complex	Protein-protein complex	Protein-nucleic acid complex	Nucleic acid-ligand complex
2004	28,991	2,276	2,276	N.A.	N.A.	N.A.
…	…	…	…	…	…	…
2018	135,859	19,588	16,151	2,416	896	2018
2019	146,836	21,382	17,679	2,594	973	136
2020	157,974	23,496	19,443	2,852	1,052	149
2021	171,254	27,408	22,920	3,176	1,141	171

*: Some earlier versions (v.2005 – v.2017) are not included in this table due to space limit.

The hierarchical structure of the PDBbind data set

The data sets in PDBbind v.2021 are compiled through a stepwise process as follows.



Protein Data Bank
171,254

Valid complexes
85,829

The general sets
27,408

The refined set*
TBD

*** This data set contains only complexes formed between proteins and small-molecule ligands.**

A. The PDBbind v.2021 is based on the contents of PDB released at the first week of year 2021, which contained a total of 171,254 experimentally determined structures.

B. All PDB structures are analyzed to identify four major categories of molecular complexes, including protein-small ligand, nucleic acid-small ligand, protein-nucleic acid and protein-protein complexes. This step identifies a total of 85,829 PDB entries as valid complexes.

C. Relevant publications are then examined to curate experimentally determined binding affinity data (K_d, K_i or IC_{50}) for all valid complexes. Binding data for 27,408 complexes have been collected in this way. They form the main body of the PDBbind database, which is referred to as the "**general set**".

D. An additional "**refined set**" is compiled to select the protein-ligand complexes with better quality out of the general set. A number of filters regarding binding data, crystal structures, and other features are applied to sample selection (check ref.1 below for details). At this moment, the refined set in PDBbind v.2021 is yet to be compiled because we plan to re-define the filters applied to the selection of this data set.

Policy for Registration and Subscription

It is FREE to register on the PDBbind+ web site to become **a demo user**. Demo users may access the contents of PDBbind up to version 2020. However, to gain full access to the latest contents of PDBbind, e.g. version 2021, as well as the useful functions implemented on PDBbind+, one needs to become **a subscriber** by paying a certain amount of license fee. Please check the information about this matter on the PDBbind+ web site (http://www.pdbbind-plus.org.cn/). For the sake of current PDBbind users, the old web site PDBbind-CN (www.pdbbind.org.cn) will be up-running as is, but no future update of PDBbind-CN is planned.

PART 2. Structure Files in the PDBbind Data Package

What is the PDBbind “general set”?

The PDBbind database covers four major types of molecular complexes in the Protein Data Bank. Those complexes with known experimental binding data (K_d, K_i or IC_{50} values) form the so-called “general set”. PDBbind version 2021 provides binding data for a total of 27,408 molecular complexes, including protein-ligand complexes (22920), nucleic acid-ligand complexes (171), protein-nucleic acid complexes (1141), and protein-protein complexes (3176).

Because the protein-ligand complex data set is the largest and by far the most popular one, currently **the “general set” by default refers to the protein-ligand complex data set**. As a valuable feature of PDBbind, we provide processed structure files for all protein-ligand complexes in the general set. Those “clean” structure files can be readily utilized by most molecular modeling software and thus bring great convenience to the users.

How are the structure files presented in the PDBbind data package?

The original PDB structure of each protein-ligand complex in the general set is processed, and the resulting files are saved in a folder named after the PDB code:

e.g. **1bxo/**

The complex structure is split into a protein molecule saved in the PDB format and a ligand molecule saved in the Tripos Mol2 format and the MDL SDF format:

e.g. **1bxo_protein.pdb**, **1bxo_ligand.mol2** & **1bxo_ligand.sdf**

For convenience in display or analysis, another PDB file is provided that includes only the binding pocket, , i.e. all residues within 10A from the ligand, on the protein molecule:

e.g. **1bxo_pocket.pdb**

For the users' convenience, a number of index files are provided, which summarize the basic contents of the PDBbind data sets. Those index files can be found under the "**index/**" folder in the PDBbind data package, including:

- "**INDEX_general_PL.2021**": List of the protein-ligand complexes with known binding data, i.e. the “general set” of protein-ligand complexes.
- "**INDEX_general_PL_data.2021**": List of the general set of protein-ligand complexes with formatted binding data.
- "**INDEX_general_PL_name.2021**": List of the general set of protein-ligand complexes with protein names and UniProt IDs.
- "**INDEX_general_PN.2021**": List of the protein-nucleic acid complexes with known binding data.
- "**INDEX_general_PP.2021**": List of the protein-protein complexes with known binding data.
- "**INDEX_general_NL.2021**": List of the nucleic acid-ligand complexes with known binding data.

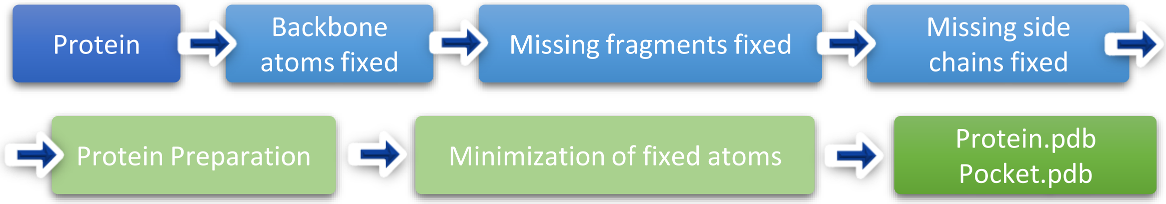
More details on processing the protein-ligand complex structures

Details about how the protein-ligand complex structures are processed will be described in our future publication. Here are a few issues that need to be particularly mentioned:

- **First of all, the protein-ligand complex structure is split into a protein molecule and a ligand molecule.** Here, the protein molecule typically contains a complete "biological unit" because it is assumed to match the binding data better. Sometimes binding of the ligand molecule involves multiple biological units, and thus in such a case, the protein molecule contains all relevant biological units. The biological unit of each complex is also downloaded from PDB.

  Coordinates of both the protein molecule and the ligand molecule generally remain as those in the original PDB structure. In other words, their coordinates are not adjusted, for example, by energy minimization. For the protein molecule, all its atoms are re-numbered continuously starting from 1 because some molecular modeling software require so. But the residue numbers and chain labels remain the same as those in the original PDB structure.

- **Structural defects on the protein molecule are fixed:** A significant number of PDB structures contain certain defects on the protein molecule, such as missing backbone atoms, residue side chains, short or long loop regions. As a new feature in PDBbind version 2021, we have attempted to fix those structural defects by filling up the missing atoms appropriately with some computer software. Those software still fail in some cases, but a large percentage of those structural defects can be fixed. In order to label the atoms added by us, the B-factor of such an atom is set to "999.99" in the ATOM/HETATM section in the PDB-format file.


Protein → Backbone atoms fixed → Missing fragments fixed → Missing side chains fixed → Protein Preparation → Minimization of fixed atoms → Protein.pdb Pocket.pdb

- **Other components are preserved with the protein structure:** Aside from the ligand molecule, other components in the original PDB structure, such as metal ions and water molecules, are saved with the protein molecule in the "HETATM" section of the final PDB-format file. In particular, saccharides and some cofactors are often observed as attachments to the main protein structure. In previous versions of the PDBbind database, except for metal ions and water molecules, other attachments to the protein molecule are completely removed. Starting from version 2021, components covalently bound to the main protein structure are also preserved in the HETATM section of the PDB-format file to keep the integrity of the original PDB structure.

- **Boundary between peptide and protein**: A frequently asked question by the PDBbind users is how we differentiate a "peptide" binder from a "protein" binder because the former is a protein-ligand complex while the latter is a protein-protein complex. By our definition, a valid protein-protein complex should consist of at least two different protein molecules, each of which should contain at least 20 residues. If the binder peptide chain is shorter than 20 residues, it is considered as a peptide ligand.

To continue on the next page

- **Interpretation and processing of the ligand molecule:** The chemical structure of each ligand molecule is interpreted with a set computer program based on the original PDB structure. Since version 2021, a completely new workflow is adopted for this purpose, where the ligand structure is first processed and saved in a SDF-format file, and then further converted into a Mol2-format file. All resulting structures are examined manually to correct atom/bond types if necessary.
- **Processing the covalently bound ligands:** In version 2021, special efforts are made to process covalently bound ligands more appropriately because it could be conceptually controversial how to split a covalent protein-ligand complex. We correct the ligand structure according to the type of the covalent bond formed with the protein. For example, if there are a leaving group on the ligand after covalent binding, the corresponding part is copied from the protein structure and added back to the ligand structure. In any circumstance, the coordinates of the ligand molecule from the original PDB structure are not adjusted.
- **Setting the protonation state:** As a new feature in PDBbind version 2021, the protonation state of the chemical groups on the ligand molecule under the neutral pH condition is determined by using *Epik* in the Schrödinger software. The determined protonation states are applied to both the SDF-format file and the Mol2-format file. At the protein molecule side, the protonation state is set by using *ProtAssign* in the *Prepwizard* module in the Schrödinger software. In addition, "flipped side chains" are allowed for His, Asn and Gln residues during the process of protein structures.

Important: New workflow for preparing the general set version 2021

To prepare PDBbind version 2021, we have thoroughly re-designed the workflow for processing the protein-ligand complex structures sourced from PDB. Our new workflow attempts to prepare both the ligand and the protein structures properly while rectifying miscellaneous issues within them. This new workflow is applied to processing the protein-ligand complex structures newly added to version 2021. It is also applied to processing the complex structures contained in the previous version and has resolved numerous miscellaneous problems accumulated in the past.

Moreover, we have adopted a higher standard for compiling the general set. Protein-ligand complexes with significant problems either in structure (e.g. ligand with poor completeness) or binding data (e.g. ambiguous data like IC_{50}>1000μM) are not accepted into this data set. This leads to the elimination of several hundred of protein-ligand complexes in the previous general set (version 2020). Details about the new filters will be revealed in our future publication.

Consequently, we are confident that besides the increase in data size, **the overall quality of the protein-ligand complexes in PDBbind version 2021 is considerably improved** as compared to the previous version. This will pave a solid ground for developing new computational or machine-learning models by using the PDBbind database.

Latest update: Feb 2024