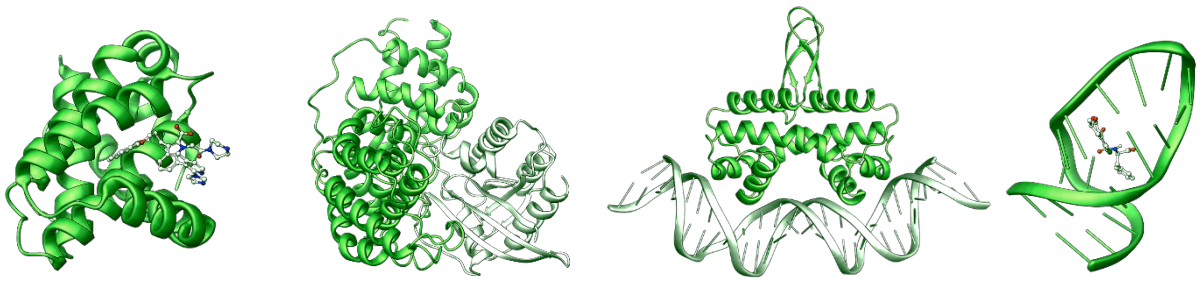
# A Brief Introduction to the PDBbind Database





Protein-ligand complex | Protein-protein complex | Protein-nucleic acid complex | Nucleic acid-ligand complex

## What is the PDBbind database? What about version 2021?

The PDBbind database provides a systematic collection of experimental binding affinity data for all types of biomolecular complexes in the Protein Data Bank (PDB). It provides an essential linkage between complex structures and their binding data, which is much needed by various computational and machine-learning studies on molecular interaction. A 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 of PDBbind is version 2021**. This release provides binding data of a total of 27408 biomolecular complexes, including protein-ligand (22920), nucleic acid-ligand (171), protein-nucleic acid (1141), and protein-protein complexes (3176).

| 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.

## What does the PDBbind database provide?

- **Carefully curated 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 ready for use:** 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 platform supported by cloud computing:** Users can access PDBbind through a web portal called PDBbind+ (http://www.pdbbind-plus.org.cn/). On this web site, basic information of each complex is displayed on a single page, and text- and structure-based search among the contents of the PDBbind data sets is enabled. This web site also incorporates valuable functional modules that are helpful for drug discovey. On-line job computation are supported by powerful cloud computing resources.

## Basic information of the PDBbind data sets

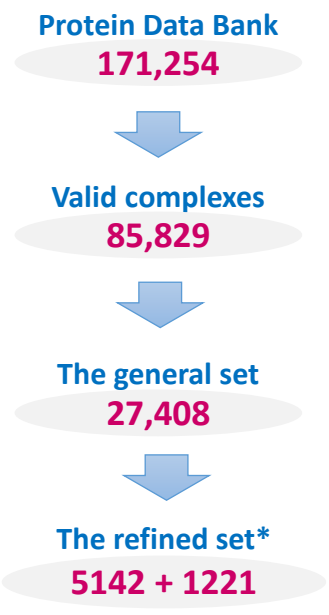
The data sets in PDBbind v.2021 are compiled through a stepwise process as follows.

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, i.e. the "**general set**".

D. An additional "**refined set**" is compiled to select the protein-ligand complexes of better quality out of the general set. A number of filters regarding binding data, crystal structures, and other features are applied to sample selection. Starting from v.2021, we provide a "standard refined set" (5142 complexes) and a "refined set plus" (5142 regular complexes plus 1221 complexes formed by metal-containing proteins and their ligands).




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

## About the structure files included in the data package

The PDBbind database covers four major types of molecular complexes in the Protein Data Bank, i.e. protein-ligand complexes, nucleic acid-ligand complexes, protein-nucleic acid complexes, and protein-protein complexes. Since the protein-ligand complex data set is the largest and by far the most popular one, as a valuable feature of PDBbind, we provide processed structure files for all protein-ligand complexes in the general set. These "clean" structure files can be readily utilized by most molecular modeling software and thus bring great convenience to the users.

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 its PDB code, in the PDBbind data package:

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 on the protein, i.e. all residues within 10A from the ligand:

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.

| File Name | Description |
|---|---|
| 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_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. |
| INDEX_refined_PL.2021 | List of the protein-ligand complexes in the "standard refined set". |
| INDEX_refined+_PL.2021 | List of the extra protein-ligand complexes in the "refined set plus". which are formed by metal-containing proteins and their ligands. |
| INDEX_structure.2021 | List of the protein-ligand complexes with processed structure for download. Currently, this list equals to the general set. |

## How are the protein-ligand complex structures processed?

### ■ New workflow for processing the protein-ligand complex structures in version 2021

For PDBbind version 2021, we have thoroughly redesigned the workflow for processing protein-ligand complex structures sourced from the PDB. Our new workflow aims to properly prepare both the ligand and protein structures while addressing various issues within them. This updated process has been applied to the newly added protein-ligand complex structures in version 2021 and to those in previous versions, resolving many problems accumulated in the past.

Moreover, we have implemented stricter criteria for compiling the general set. Protein-ligand complexes with significant structural issues (e.g., largely incomplete ligands) or problematic binding data (e.g., ambiguous data like $IC_{50}$ >1000μM) are excluded from this dataset. As a result, several hundred protein-ligand complexes from the previous general set have been eliminated.

As a result, we are confident that, in addition to the increased data size, the overall quality of the protein-ligand complexes in PDBbind version 2021 has significantly improved compared to the previous version. It will represent a more solid foundation for developing new computational or machine-learning models using the PDBbind database.

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, 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.


*To continue on the next page*

- **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.
- **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.

## Wishlist

Protein-ligand complexes | Protein-protein complexes | Protein-nucleic acid complexes | Nucleic acid-ligand complexes

Currently, we provide processed structural files only for complexes formed by proteins and small-molecule ligands. Processed structural files for other types of complexes, such as protein-protein or protein-nucleic acid complexes, are not yet available. However, we plan to cover these additional types of complexes as well in forthcoming versions.

# Description of the "refined set"

## ■ What is the refined set and why do we need it?

As an added value of PDBbind, a "refined set" is selected from the protein-ligand complexes in the general set by filtering out complexes with various structural, binding data, or biological/chemical issues. In fact, only one-fourth of the protein-ligand complexes in the general set are included in the refined set, reflecting the higher quality of this data set. The refined set thus provides a more robust choice for docking/scoring and many other types of research.

In particular, with the introduction of a new workflow for processing complex structures in version 2021, the overall quality of the protein-ligand complexes in the general set—and consequently the refined set—has improved significantly. Many accumulated issues from previous versions have been resolved, and more stringent criteria are applied to selection of new samples into the refined set. As a result, the refined set version 2021 has reached a new level in both size and quality.

## ■ How is the refined set selected?

A number of filtering rules are applied to sample selection, which are summarized as follows:

**Category I: Concerns on the quality of the complex structure**

- Only complexes with crystal structures are accepted; NMR-resolved structures are not.
- Resolution of the complex structure must be better than 2.5 Å, and R-factor < 0.250.
- The ligand structure in the complex must be complete, without missing atoms/fragments.
- If there are any missing segment of the backbone in the binding pocket (within 10 Å from the ligand) on the protein structure, the complex is not accepted.
- Covalent complexes are not accepted.
- A non-covalent complex is not accepted if any significant steric clash (distance < 2.0 Å) exists between a heavy atom pair between the protein and the ligand.

**Category II: Concerns on the quality of the binding data**

- Complexes with known dissociation constants ($K_d$) or inhibition constants ($K_i$) are accepted. Complexes with only half-inhibition or half-effect concentrations ($IC_{50}$ or $EC_{50}$) values are not.
- Complexes with extremely low ($K_d$ or $K_i$ > 10 mM) or extremely high ($K_d$ or $K_i$ < 1 pM) binding affinities are not accepted.
- Estimated binding data, e.g. "$K_d$ ~ 1 nM" or "$K_i$ > 10 μM", are not accepted.
- If the protein molecule has multiple binding sites which are associated with significantly different binding constants (> 10 folds), this complex is not accepted.
- The protein molecule used in binding assay has to match the one used in crystal growth, i.e. the same species, subtype, and mutation. Similarly, the ligand molecule used in binding assay has to match the one used in crystal growth.

*To continue on the next page*

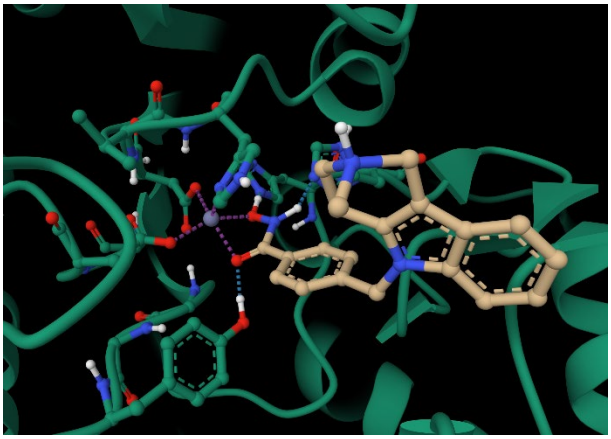
**Category III: Concerns on the biological/chemical features of the complex**

- Molecular weight of the ligand molecule must be lower than 1000 if it is a regular organic molecule. It must not contain more than 10 residues if it is a poly-peptide, or it must not contain more than 3 residues if it is a poly-nucleotide.
- The ligand molecule must not consist of atoms other than carbon, nitrogen, oxygen, phosphorus, sulfur, halogen and hydrogen atoms.
- If the buried surface area of the ligand molecule is below 33% of its total surface area, this complex is not accepted.
- The binding site on the protein molecule must not contain any non-natural amino acid residues in direct contact with the bound ligand (distance < 5 Å).
- Only binary complexes are accepted, i.e. the complex must be formed distinctly between one protein molecule and one ligand molecule. Ternary complexes are not accepted, e.g. a cofactor bound closely to the ligand (distance < 5 Å) at the same binding site.
- If the complex is essentially redundant to another complex, i.e. the same protein and the same ligand) but at a lower resolution, it is not accepted.

## ■ The standard refined set and the refined set plus

Starting with version 2021, we offer a "**standard refined set**" containing 5,142 complexes and a "**refined set plus**", which includes the same 5,142 complexes plus 1,221 complexes. The extra 1,221 complexes in the refined set plus meet the same quality standards as those in the standard refined set but are formed by metal-containing proteins and their ligands. Users can find a folder named **"refined+"** in the PDBbind data package, where the processed structural files for all protein-ligand complexes in the refined set plus are assembled.

It is worth noting that all previous versions of the refined set included complexes formed by metal-containing proteins. However, we have observed that many PDBbind users tend to remove these complexes when training or testing their models or for other purposes. We assume these users believe that such complexes are more similar to covalent complexes and should therefore be excluded. This perspective is, in fact, debatable: while the bond between a metal atom and a ligand molecule is partially covalent, it is also reversible, distinguishing it from a typical covalent bond.



**Histone deacetylase 10 in complex with tubastatin A (PDB code 6WBQ)**

Anyway, we have now separated the complexes formed by metal-containing proteins from the regular protein-ligand complexes. Technically, this separation is straightforward. With this change, the standard refined set and the refined set plus can now be used more flexibly to suit different user needs. It is now up to the user to decide whether to combine or separate these two data sets.


## Policy for registration and subscription

We have taken a significant step towards the commercialization of the PDBbind database starting from version 2021. The PDBbind+ web site serves as the exclusive gateway to both the current and forthcoming releases of the PDBbind database. Full access to the PDBbind database and the valuable features integrated into the PDBbind+ web site is reserved for **paid subscribers**. Please check the information about this matter on the PDBbind+ web site (www.pdbbind-plus.org.cn).

Nevertheless, it is FREE to register on the PDBbind+ web site. Upon completion of registration, the user is automatically designated as a **demo user**, providing the user limited access to the available features on PDBbind+, for example, access to the contents of the PDBbind database up to version 2020.

The old PDBbind-CN web site (www.pdbbind.org.cn) , which has hosted the PDBbind database up to version 2020, will still be up-running as is, but no future update of PDBbind-CN is planned.

## References and Notes

The PDBbind database is developed by Prof. Renxiao Wang's group at the School of Pharmacy, Fudan University in Shanghai, People's Republic of 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)

*Latest update: Aug 2024*

For technical issues, please contact us at *support@pdbbind-plus.org.cn*
For sale issues, please contact us at *sales@pdbbind-plus.org.cn*