BioLiP2: a database for biological unit-based protein-ligand interactions

Hong Wei [1], Wenkai Wang [1], Zhenling Peng [2, *], and Jianyi Yang [2, *]

[1] School of Mathematical Sciences, Nankai University, Tianjin 300071, China

[2] MOE Frontiers Science Center for Nonlinear Expectations, Research Center for Mathematics and Interdisciplinary Sciences, Shandong University, Qingdao 266237, China

* To whom correspondence should be addressed, e-mail: yangjy@sdu.edu.cn or zhenling@email.sdu.edu.cn

ABSTRACT

Since its establishment in 2013, BioLiP has become one of the widely used databases for protein-ligand interactions. Nevertheless, several key issues occurred with it over the past decade. For example, the adopted format is single chains for both proteins and ligands, which is inappropriate as many interactions are formed with multiple protein chains (known as biological units) and multiple ligand chains (e.g., DNA). We sought to address these issues, resulting in BioLiP2, a new database for protein-ligand interactions. The major features of BioLiP2 include: (1) protein structures are in the form of biological units rather than single-chain structures; (2) DNA/RNA chains are properly paired rather than separated; (3) both experimental and predicted binding affinities are provided; (4) both biologically relevant and irrelevant protein-ligand interactions are retained; (5) a new biological unit structure-based ligand-binding site prediction algorithm is developed. With these new features, BioLiP2 is expected to be a valuable resource for studying protein-ligand interactions. It is freely available at: http://yanglab.qd.sdu.edu.cn/BioLiP2/.

INTRODUCTION

The biological function of many proteins, known as receptors, is achieved by interacting with small molecules, which are referred to as ligands. To enable related computational studies, such as the prediction of the protein-ligand binding site (1-3), prediction of binding affinity (4-7), protein-ligand docking (8-11), and structure-based drug design (12-14), etc, a collection of high-quality protein-ligand interaction data is essential (15-18). BioLiP is a database that collects 3D structures of biologically relevant protein-ligand interactions (19). It has emerged as one of the most widely used resources for investigating protein-ligand interactions.

There are several known inherent issues with the data in BioLiP. The first one is that the receptor structure is presented in a single-chain form. Nonetheless, the functional form of many proteins, called biological units, typically comprises multiple interacting chains. Consequently, some vital protein-ligand interactions are not captured in the BioLiP data due to the incompleteness of the receptor structure (20), particularly when a ligand simultaneously interacts with multiple protein chains. For example, the hemoglobin protein can only transport oxygen in the form of a tetramer, which is composed of four chains. Second, DNA ligands in BioLiP are presented in a single-chain format, which is inconsistent with the fact that DNAs typically form double-helix structures consisting of two complementary chains. The third one is the potential misjudgement of the biological relevance. BioLiP employs an empirical rule to determine whether a protein-ligand interaction is biologically relevant or not. Consequently, a protein-ligand interaction that is deemed biologically irrelevant is excluded from BioLiP, resulting in missing data when a biologically relevant interaction is mistakenly considered irrelevant.

In this article, we introduce an improved version of BioLiP, named BioLiP2, which aims to address the aforementioned issues. The major updates include the representation of receptor structures in biological units; the retention of both biologically relevant and irrelevant interactions; the paired chain structures for DNA/RNA ligands; the adoption of mmCIF format (21) rather than PDB format to deal with huge molecules (that exceed the capability of the PDB format); last but not the least, we provide an efficient template-based approach to ligand-binding site prediction using the BioLiP2 data, which allows for the input with either single-chain or multiple-chain receptor structures.

MATERIAL AND METHODS

Overview of BioLiP2

The overview flowchart for building BioLiP2 is presented in Figure 1. First, starting from the asymmetric unit files (in mmCIF format), biological units are generated using the rotation and translation matrices stored in the mmCIF files. Then, receptors and ligands are extracted from the biological units. In this work, receptors are defined as proteins with at least 30 amino acids in each polypeptide chain while other molecules are defined as ligands (including small molecules, peptides with less than 30 amino acids, and DNA/RNA). As base pairs can be formed between DNA/RNA chains, a heuristic algorithm (see "**An effective DNA/RNA pairing algorithm**") is proposed to pair the DNA/RNA chains. As done in BioLiP (19), the biological relevance of each ligand-protein complex is assessed with a semi-manual procedure. For the sake of completeness, BioLiP2 also retains biological units with ligands or receptors only. Thus, the data deposited in BioLiP2 consist of three categories: biologically relevant ligand-protein interactions, biologically irrelevant ligand-protein interactions, and structures with ligands or proteins only. They are accessible in a user-friendly web interface at our website.

Procedure for database construction

The raw data are downloaded from the PDB database (22). The procedure for the construction of BioLiP2 consists of three major steps.

Step 1. Generation of biological units. For each PDB entry, the asymmetric unit (in mmCIF format) is first downloaded. One or more biological units are generated according to the rotation matrices and the translation vectors stored in the mmCIF file's records, i.e., 'pdbx_struct_assembly', 'pdbx_struct_assembly_gen', and 'pdbx_struct_oper_list'. Modified residues are converted to standard ones based on the record 'pdbx_struct_mod_residue'.

Step 2. Extraction of receptor and ligands from biological units. Each biological unit is separated into one receptor and multiple ligands (when exist). The receptor consists of protein chains with at least 30 amino acids; while ligands can be small molecules, DNA/RNA, or peptides with less than 30 amino acids.

Step 3. Annotations of protein-ligand interactions. The interaction between the receptor and each ligand is annotated in terms of binding residues, biological relevance, and binding affinity.

(1) *Binding residues*. Binding residues are determined based on the atomic distances between the protein and ligand atoms. A residue is considered a binding residue if its closest atomic distance to the ligand is less than a defined threshold, which is the sum of the Van der Waal's radius of the corresponding atoms plus 0.5 Å (23).

(2) *Biological relevance*. As some small molecules are used to assist the protein structure determination and thus might not perform biological functions, i.e., they are biologically irrelevant. Peptides and nucleic acids are regarded as biologically relevant to receptors without the need for assessment. We adopt the same semi-manual procedure used in BioLiP to assess the biological relevance of the small molecules. Briefly, a list of 465 small molecules that are commonly used in protein structure determination is collected manually (available at the *Download* page). Small molecules outside this list are regarded as biologically relevant by default. For molecules in this list, a hierarchical procedure is employed, which involves computing of the number and distribution of binding residues, text mining of the PubMed abstract, and manual verification of suspicious entries. More details can be found in the BioLiP paper.

(3) *Binding affinity*. The strength of ligand-protein interaction is commonly referred to as binding affinity. Experimental ligand-binding affinity is collected from three databases: Binding MOAD (16), PDBbind (17), and BindingDB (18). However, only a limited portion (5.8%) of PDB entries have experimental binding affinity data. To address this issue, predicted binding affinity data are provided in BioLiP2 for entries that do not have experimental binding affinity data. The predicted binding affinity data are obtained based on a few well-established physics-based tools, including X-Score (4), ITScore (5), and AutoDock Vina (9). To generate a consensus binding affinity prediction, a regression model was trained to fit the real binding affinity by taking the three scores as input.

(4) *Area of binding surface*. Besides binding affinity, the area of the binding interface is provided in BioLiP2. Motivated by DOCKGROUND (24), we measured the binding interface area S as below:

$$S = \frac{1}{2}\left(S_{\text{prot}} + S_{\text{lig}} - S_{\text{com}}\right) \tag{1}$$

where S_{prot}, S_{lig}, S_{com} are the solvent-accessible surface area calculated by the program FreeSASA (25), with the protein, ligand, and ligand-protein complex structures as inputs, respectively.

An effective DNA/RNA pairing algorithm

As is well known, DNA typically forms double-helix structure with two complementary chains. Inter-chain interactions are also witnessed between different DNA and RNA chains. However, in BioLiP, the DNA/RNAs are simply divided into separated chains, making the DNA/RNA structures incomplete. Thus, it is necessary to correctly pair DNA/RNA chains to keep them intact. To address this issue, we propose a simple yet effective pairing algorithm for DNA/RNA chains. To pair the DNA/RNA chains, we first parse the secondary structure of the whole DNA/RNA structure using the DSSR program (26). Then the pairing state of each nucleotide is derived from the output of DSSR. Two chains are paired together if there are at least 3 inter-chain base pairs.

RESULTS AND DISCUSSION

BioLiP2 versus BioLiP

The major differences between the data in BioLiP2 and BioLiP are summarized in Table 1, including the following five aspects.

Table 1. The major difference between the data in BioLiP2 and BioLiP.

Feature	BioLiP2	BioLiP
source data format	mmCIF	PDB
protein structure	biological unit	monomer structure
state of DNA/RNA chains	paired chains	separated chains
binding affinity	experimental + predicted	experimental only
data content	all PDB structures	biologically relevant interactions

(1) *Source data format*. Due to the limitation of the PDB format, structures with more than 62 chains such as many cryo-EM structures are ignored in BioLiP. The source data processed by BioLiP2 is in mmCIF rather than PDB format, which solves the issue of missing structures in BioLiP.

(2) *Protein structure*. One of the key developments of BioLiP2 over BioLiP is the change in the form of protein structure. BioLiP2 improves the representation of protein structure from single-chain based tertiary structure to quaternary structure (i.e., biological unit). As indicated above, the quaternary structure may consist of single chain or multiple chains. The completeness of the protein structure can ensure the completeness of ligand-binding interactions(20). For instance, Figure 2A shows the interaction

between the HIV-1 protease and its inhibitor (PDB ID: 1EBY), where the protease is a C2 symmetric homo dimer and the catalytic residues are located on the interface. This interaction is however separated into two individual entries in BioLiP (each for one receptor chain, Figure S1).

(3) *State of DNA/RNA chains*. With the pairing algorithm introduced above, we are able to correctly pair DNA/RNA chains. For example, the DNA ligand binding with the structure of an *intron encoded endonuclease* is formed by four DNA chains (PDB ID: 1A73, Figure 2B). On the contrary, BioLiP separates the interactions into 6 individual entries (see Figure S2): 2 receptor chains, and 3 DNA chains binding with each receptor chain (one DNA chain is ignored as it does not interact with the receptor chain after separation).

(4) B*inding affinity*. In BioLiP2, ~48,000 entries are annotated with experimental binding affinity data. In addition, predicted binding affinity and binding interface area are also provided for ~1.7 million protein-small molecule interactions (peptide and DNA/RNA ligands are excluded).

(5) *Data content*. Only interactions judged as biologically relevant are kept in BioLiP. However, this may result in the omission of some biologically relevant interactions due to misjudgement. For example, the structure of the *tetrameric NaK channel binding with K+* (PDB ID: 2AHZ) was judged as biologically irrelevant (due to its limited binding to single-chain structure) and thus not deposited in BioLiP. In fact, the ion K+ is located in the pocket of a tetrameric cation channel and thus biologically relevant. This is correctly captured in BioLiP2 due to the utilization of quaternary structure (Figure 2C). We admit that it is also possible that the judgement of biological relevance in BioLiP2 is not perfect. We thus keep all data in our database so that the users could make judgement based on their own expertise. For the sake of completeness, structures with ligands or proteins only are also stored in our database, which may be used for other purposes, such as in protein structure database searching. To summarize, the BioLiP2 data are organized into three categories, biologically relevant interactions; biologically irrelevant interactions; and structures with ligands or proteins only.

Statistics of the BioLiP2 data

By the time of May 17th, 2023, BioLiP2 contains about 2.6 million entries, involving about 200,000 structures from PDB. Figure 3A shows the overall distribution of these entries. Among the total entries, ~32/66% are annotated as biologically relevant/irrelevant interactions (the orange/blue region); the remaining 2% are biological units with ligands or proteins only (the black region). Further analysis is conducted on the biologically relevant interactions, which are the core data of BioLiP2.

The distribution of the proteins involved in biologically relevant interactions is shown in Figures 3B and 3C. Figure 3B suggests that almost half (48.7%) of these proteins are monomers, while the remaining half contain more than one chains (called oligomers). Interestingly, oligomers with even numbers of chains (e.g., dimmer and tetramer) are more than those with odd numbers of chains (e.g., trimer). The oligomers can be divided into two groups, homo-oligomers and hetero-oligomers, depending on if all chains are identical or not. Figure 3C shows the proportion of homo-oligomers and hetero-oligomers at different oligomeric states. For oligomers with even numbers of chains (especially for dimers and tetramer), the numbers of homo-oligomers tend to be higher than the numbers of hetero-oligomers.

Figure 3D shows the proportion of different ligand types involved in biologically relevant interactions. Regular small molecules account for more than half (52.5%) of all ligands; metal ions are the second largest group (34.3%). For nucleic acids ligands (4.4%), 61.6% and 35.5% of them are RNA and DNA, respectively; the remaining 2.9% are DNA-RNA complex. Among the biologically relevant interactions involving oligomeric targets, 30.2% of the ligands interact with more than one chains (Figure S3A), which spreads across states from dimmers to 10-mers (Figure S3B). This indicates the necessity of updating BioLiP to BioLiP2, to maintain the completeness of ligand-protein interactions.

Binding affinity data

About 48,000 entries in BioLiP2 (from ~20,000 unique PDB entries) are annotated with experimental binding affinity. Predicted binding affinities are provided for other entries without experimental binding affinity. Benchmark test on 285 high-quality data from the PDBbind core set (v2016) shows that the binding affinity predicted by these tools correlates well with the experimental binding affinity, with Person's correlation coefficient (PCC) around 0.6 (Figure S4A-C). The consensus binding affinity

prediction based on a regression model yields a 4.7~15.5% higher PCC than the individual predictions (Figure S4D). Further tests based on a bootstrap sampling on the dataset by 1000 times shows the improvement is statistically significant (Figure S5).

BioLiP2 web interface

The BioLiP2 database is freely accessible at http://yanglab.qd.sdu.edu.cn/BioLiP2/. Five modules are provided to use the data (illustrated in Figure 4), which are introduced below.

(1) *Browse module*. There are three options in the browse module: (i) Browse biologically relevant entries, (ii) Browse biologically irrelevant entries, and (iii) Browse structures with ligands or proteins only. Clicking on each option would display the summary of entries in the corresponding category in the form of a table, which has seven columns: BioLiP2 ID, PDB ID with resolution, Assembly ID, Stoichiometry, Ligand ID, Site and Binding affinity. The detailed information for each entry is available by clicking on the corresponding BioLiP2 ID.

(2) *Search module*. In the search module, users can search for specific interested data by four fields: PDB ID, UniProt ID, Ligand ID, and Ligand Name. Meanwhile, a quick search by PDB ID can be conducted on the home page. To save time, users can select to return only biologically relevant data. The result page is organized in a table. Users can choose to display more information such as UniProt ID, and PubMed ID by clicking the toggle button. An input box is provided to filter results further. The results can be downloaded in several formats including JSON, CSV, and so on.

(3) *Ligand module*. In this module, all ligands involved in BioLiP2 are displayed in two columns by default: Ligand ID and Ligand Name. More detailed information can be shown by clicking on the Ligand ID link. The detailed ligand page contains the 2D view of ligand, synonyms, as well as the chemical component summary. To sufficiently annotate ligands, the synonyms for each ligand are collected from PDB (27), ChEBI (28), and PubChem (29), which are also essential in the biological relevance assessment. The chemical component summary includes molecular weight, chemical formula, the SMILES string, and external links. Besides, a search module (the "Advanced Search" button) is provided for searching for a specific ligand by ligand ID or ligand Name.

(4) *Biological unit-based binding site prediction module COACH-Q*. COACH-Q is an extension of the COACH-D algorithm (30) for template-based ligand-binding site prediction. COACH-Q supports

inputs with both oligomers and monomers, and makes prediction using the biological unit-based templates in BioLiP2. It is about 40 times faster and 6.5% more accurate than COACH-D. More details about COACH-Q will be introduced elsewhere. Users can provide either experimental structure or predicted structure as input, the server will return the top five predictions in 20 minutes on average.

(5) *Download module.* The BioLiP2 data are provided for download at this module. There are three tables provided on the download page: Data, Script, and Weekly Update. In the Data table, both redundant and non-redundant versions are provided for easy download. The redundant version contains all interaction data (~2.5 million interactions, i.e., biologically relevant + irrelevant interactions) and the structures with ligands or proteins only. The non-redundant version is obtained by removing redundancy from biologically relevant interactions. As biological unit structure may contain multiple chains, we define redundancy as follows. First, receptor structures are grouped by the oligomeric state (i.e. the number of chains). Then within each group, two structures are considered redundant if there exists any pair of chains from the two structures exhibit a sequence identity higher than 95%. The non-redundant version has ~40,000 receptors and ~200,000 ligands. Analysis suggests that the distributions of the ligands binding to multiple chains or single chain are similar to the redundant dataset (Figure S3C, D).

Given that some users may be unfamiliar with the mmCIF format, we provide two scripts to convert from mmCIF format to PDB format, for receptor and ligand structures, respectively. BioLiP2 is updated weekly with PDB. The updated data are provided for download in the Weekly Update table.

For users who are interested in monomer structure-based protein-ligand interaction, we also provide the redundant data in single-chain format like BioLiP. They are obtained by splitting the biological units into individual receptor chains. The data are available for download at a separate link (http://yanglab.qd.sdu.edu.cn/BioLiP2/Download/index_biolip.html).

Detailed information on protein-ligand interactions

The information on protein-ligand interactions is presented as a webpage, which consists of three tabs. The first tab is the 3D visualization of the interaction (Figure 5A). The visualization is powered by the 3Dmol package (31). Two buttons are provided to show the local view and specific binding residues, respectively. In addition, the specific binding residues can be shown in a table by clicking on the "show binding residues" button. On the right of the page (Figure 5B), the overall information is organized into

three sections: *Receptor*, *Ligand*, and *Download*. In the *Receptor* section, the stoichiometry type is derived with the same definition as in PDB, i.e., two chains are considered equivalent if their sequence identity is greater than 95%. In the *Ligand* section, three items are shown: ligand ID, ligand name, as well as the biological relevance. In the *Download* section, the structures of receptor and ligand are provided in both mmCIF and PDB formats. Structures comprising more than 62 chains are currently unavailable in the legacy PDB format. In order to maintain completeness, these large structures are divided into multiple PDB files. A tarball (.tar.gz) is provided on the page, which is a bundle of these files along with a text file to map the original chain IDs to the new IDs.

The second tab is the ligand information (Figure 5C), including ligand ID (with a link to the detailed information page), ligand name, synonyms, formula, and 2D image of the ligand structure.

The last tab is the binding information (Figure 5D), which consists of the binding affinity and the surface area. Both experimental (if available) and predicted binding affinities are provided.

CONCLUSION

We have developed the BioLiP2 database for biological unit-based protein-ligand interactions. The major contributions include: (1) All structures in BioLiP2 are based on biological units, making the protein-ligand interactions complete; (2) DNA/RNA chains are properly paired; (3) Both experimental and predicted binding affinities are provided; (4) To address the problem of misjudgement of biological relevance, both relevant and irrelevant entries are kept; (5) A biological unit-based ligand-binding site prediction algorithm is provided. We believe that the development of BioLiP2 will better serve the community of protein-ligand interactions and facilitate modern drug discovery.

AVAILABILITY

BioLiP2 is freely available at: http://yanglab.qd.sdu.edu.cn/BioLiP2/.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR online.

ACKNOWLEDGEMENT

This work is supported in part by the National Natural Science Foundation of China (NSFC T2225007, T2222012). Funding for open access charge: NSFC.

CONFLICT OF INTEREST

None.

REFERENCES

1. Yang, J., Roy, A. and Zhang, Y.J.B. (2013) Protein–ligand binding site recognition using complementary binding-specific substructure comparison and sequence profile alignment. **29**, 2588-2595.
2. Wu, Q., Peng, Z., Zhang, Y. and Yang, J.J.N.a.r. (2018) COACH-D: improved protein–ligand binding sites prediction with refined ligand-binding poses through molecular docking. **46**, W438-W442.
3. Laurie, A.T. and Jackson, R.M.J.B. (2005) Q-SiteFinder: an energy-based method for the prediction of protein–ligand binding sites. **21**, 1908-1916.
4. Wang, R., Lai, L. and Wang, S. (2002) Further development and validation of empirical scoring functions for structure-based binding affinity prediction. *Journal of computer-aided molecular design*, **16**, 11-26.
5. Huang, S.Y. and Zou, X. (2006) An iterative knowledge-based scoring function to predict protein-ligand interactions: II. Validation of the scoring function. *Journal of computational chemistry*, **27**, 1876-1882.
6. Jiménez, J., Skalic, M., Martinez-Rosell, G., De Fabritiis, G.J.J.o.c.i. and modeling. (2018) K deep: protein–ligand absolute binding affinity prediction via 3d-convolutional neural networks. **58**, 287-296.
7. Ballester, P.J. and Mitchell, J.B.J.B. (2010) A machine learning approach to predicting protein–ligand binding affinity with applications to molecular docking. **26**, 1169-1175.
8. Lemmon, G. and Meiler, J. (2012) Rosetta Ligand docking with flexible XML protocols. *Methods in molecular biology (Clifton, N.J.)*, **819**, 143-155.
9. Trott, O. and Olson, A.J. (2010) AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *Journal of computational chemistry*, **31**, 455-461.
10. Liu, Y., Yang, X., Gan, J., Chen, S., Xiao, Z.-X. and Cao, Y.J.N.A.R. (2022) CB-Dock2: Improved protein–ligand blind docking by integrating cavity detection, docking and homologous template fitting. **50**, W159-W164.
11. Liu, Y., Yang, X., Gan, J., Chen, S., Xiao, Z.X. and Cao, Y. (2022) CB-Dock2: improved protein-ligand blind docking by integrating cavity detection, docking and homologous template fitting. *Nucleic Acids Res*, **50**, W159-W164.
12. Vriend, G. (1990) WHAT IF: A molecular modeling and drug design program. *Journal of Molecular Graphics*, **8**, 52-56.

13. Anderson, A.C.J.C. and biology. (2003) The process of structure-based drug design. **10**, 787-797.
14. Kuntz, I.D.J.S. (1992) Structure-based strategies for drug design and discovery. **257**, 1078-1082.
15. Maietta, P., Lopez, G., Carro, A., Pingilley, B.J., Leon, L.G., Valencia, A. and Tress, M.L. (2014) FireDB: a compendium of biological and pharmacologically relevant ligands. *Nucleic Acids Res*, **42**, D267-272.
16. Benson, M.L., Smith, R.D., Khazanov, N.A., Dimcheff, B., Beaver, J., Dresslar, P., Nerothin, J. and Carlson, H.A. (2007) Binding MOAD, a high-quality protein–ligand database. *Nucleic acids research*, **36**, D674-D678.
17. Wang, R., Fang, X., Lu, Y., Yang, C.-Y. and Wang, S. (2005) The PDBbind Database: Methodologies and Updates. *Journal of Medicinal Chemistry*, **48**, 4111-4119.
18. Gilson, M.K., Liu, T., Baitaluk, M., Nicola, G., Hwang, L. and Chong, J. (2015) BindingDB in 2015: A public database for medicinal chemistry, computational chemistry and systems pharmacology. *Nucleic acids research*, **44**, D1045-D1053.
19. Yang, J., Roy, A. and Zhang, Y. (2013) BioLiP: a semi-manually curated database for biologically relevant ligand-protein interactions. *Nucleic acids research*, **41**, D1096-1103.
20. Dai, Y., Li, Y., Wang, L., Peng, Z. and Yang, J. (2022) On Monomeric and Multimeric Structures-Based Protein-Ligand Interactions. *IEEE/ACM transactions on computational biology and bioinformatics*, **19**, 569-574.
21. Bourne, P.E., Berman, H.M., McMahon, B., Watenpaugh, K.D., Westbrook, J.D. and Fitzgerald, P.M. (1997) Macromolecular Crystallographic Information File. *Methods Enzymol*, **277**, 571-590.
22. Rose, P.W., Beran, B., Bi, C., Bluhm, W.F., Dimitropoulos, D., Goodsell, D.S., Prlic, A., Quesada, M., Quinn, G.B., Westbrook, J.D. *et al.* (2011) The RCSB Protein Data Bank: redesigned web site and web services. *Nucleic acids research*, **39**, D392-401.
23. Schmidt, T., Haas, J., Gallo Cassarino, T. and Schwede, T. (2011) Assessment of ligand-binding residue predictions in CASP9. *Proteins*, **79 Suppl 10**, 126-136.
24. Kundrotas, P.J., Anishchenko, I., Dauzhenka, T., Kotthoff, I., Mnevets, D., Copeland, M.M. and Vakser, I.A.J.P.S. (2018) Dockground: a comprehensive data resource for modeling of protein complexes. **27**, 172-181.
25. Mitternacht, S. (2016) FreeSASA: An open source C library for solvent accessible surface area calculations. *F1000Research*, **5**, 189.
26. Lu, X.-J., Bussemaker, H.J. and Olson, W.K. (2015) DSSR: an integrated software tool for dissecting the spatial structure of RNA. *Nucleic acids research*, **43**, e142-e142.
27. Rose, P.W., Beran, B., Bi, C., Bluhm, W.F., Dimitropoulos, D., Goodsell, D.S., Prlić, A., Quesada, M., Quinn, G.B. and Westbrook, J.D.J.N.a.r. (2010) The RCSB Protein Data Bank: redesigned web site and web services. **39**, D392-D401.
28. Hastings, J., Owen, G., Dekker, A., Ennis, M., Kale, N., Muthukrishnan, V., Turner, S., Swainston, N., Mendes, P. and Steinbeck, C. (2016) ChEBI in 2016: Improved services and an expanding collection of metabolites. *Nucleic acids research*, **44**, D1214-1219.

29. Kim, S., Chen, J., Cheng, T., Gindulyte, A., He, J., He, S., Li, Q., Shoemaker, B.A., Thiessen, P.A., Yu, B. *et al.* (2020) PubChem in 2021: new data content and improved web interfaces. *Nucleic acids research*, **49**, D1388-D1395.
30. Wu, Q., Peng, Z., Zhang, Y. and Yang, J. (2018) COACH-D: improved protein-ligand binding sites prediction with refined ligand-binding poses through molecular docking. *Nucleic acids research*, **46**, W438-w442.
31. Rego, N. and Koes, D.J.B. (2015) 3Dmol. js: molecular visualization with WebGL. **31**, 1322-1324.

FIGURES

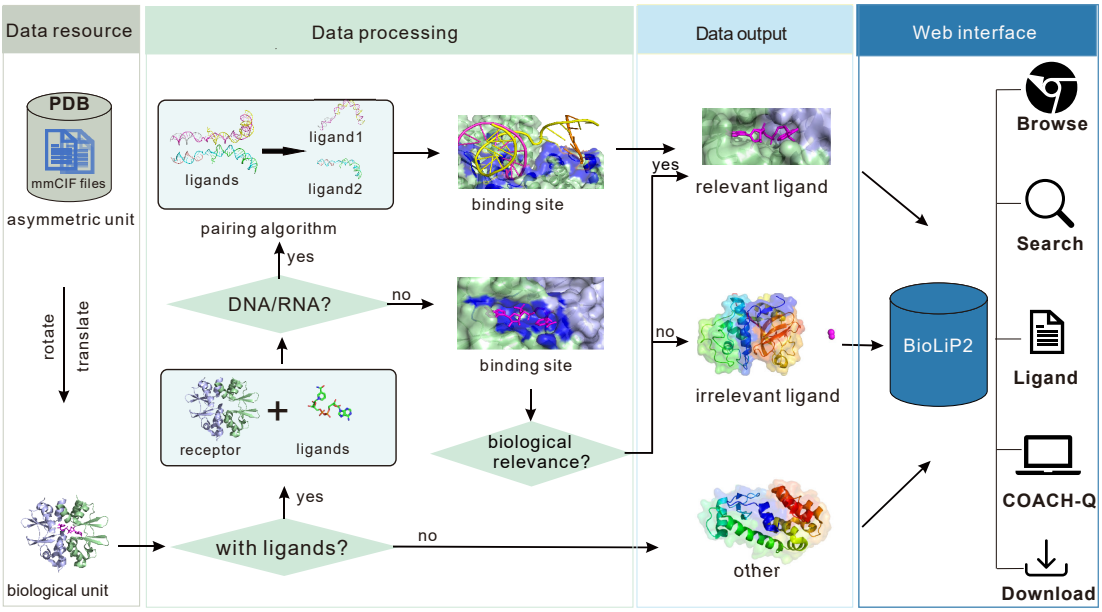

Figure 1. Workflow of BioLiP2. Biological units are first generated from the asymmetric unit files in PDB. The biological units are then processed into three categories, biologically relevant interactions, biologically irrelevant interactions, and structures with ligands or proteins only. All the data are accessible through a web interface.

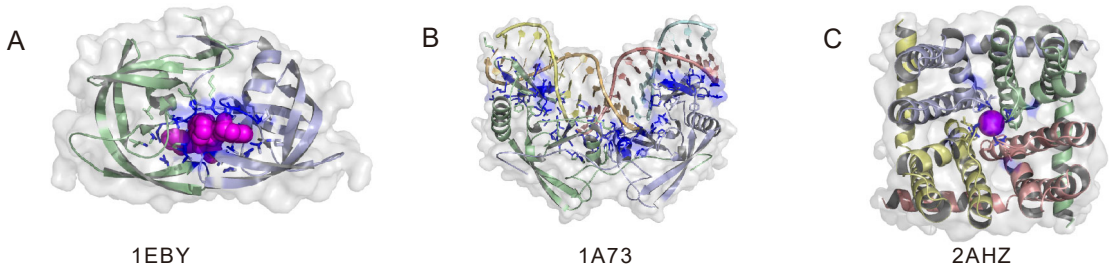

Figure 2. Three examples of protein-ligand interactions in BioLiP2. (A) The interaction between the *HIV-1 protease* and its inhibitor (PDB ID: 1EBY). (B)The DNA ligand binding with the structure of an *intron encoded endonuclease* (PDB ID: 1A73). (C) The structure of *the tetrameric NaK chanel binding with K+* (PDB ID: 2AHZ).

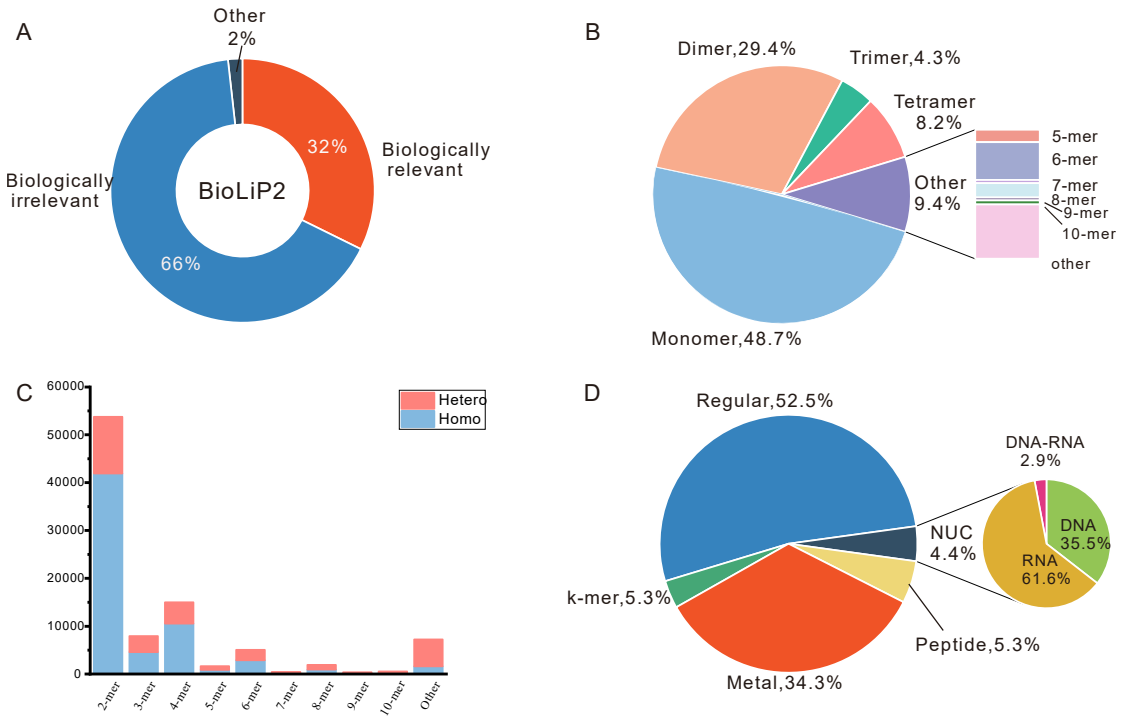

Figure 3. Statistics of the BioLiP2 data. (A) overall statistics for the BioLiP2 entries. (B) distribution of receptors in terms of the oligomeric state. (C) distribution of receptors in terms of homo- and hetero-oligomers at different oligomeric states. (D) distribution of ligands. Note that the data B-D only involve the biologically relevant interactions.

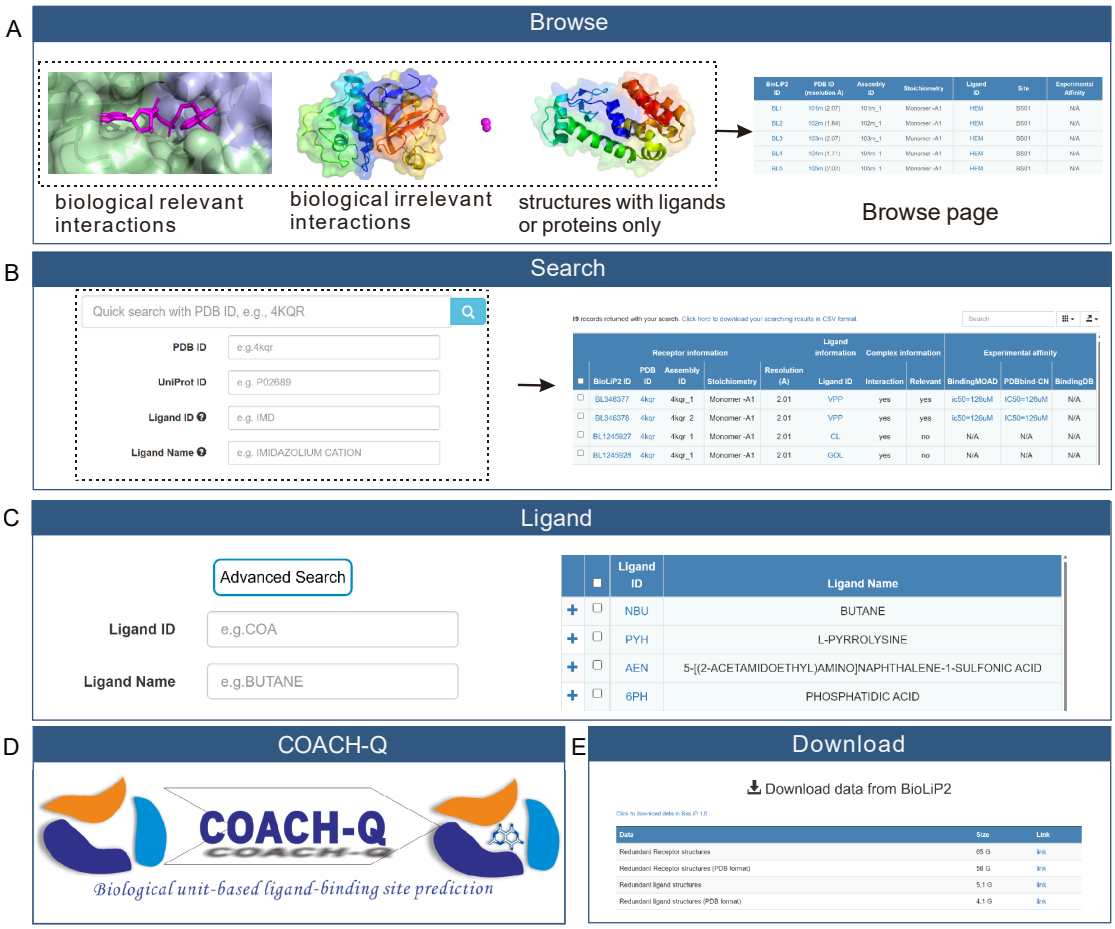

Figure 4. Major modules of the BioLiP2 web interface. (A) Through the *Browse* module, users can browse entries of three categories. (B) Through the *Search* module, users can perform search with four fields. (C) Through the *Ligand* module, users can search for detailed information for the specific ligand or browse all ligands. (D) A biological unit-based binding site prediction module *COACH-Q* is provided, which supports structure inputs with both monomers and oligomers. (E) Through the *Download* page, all data in BioLiP2 can be downloaded freely.

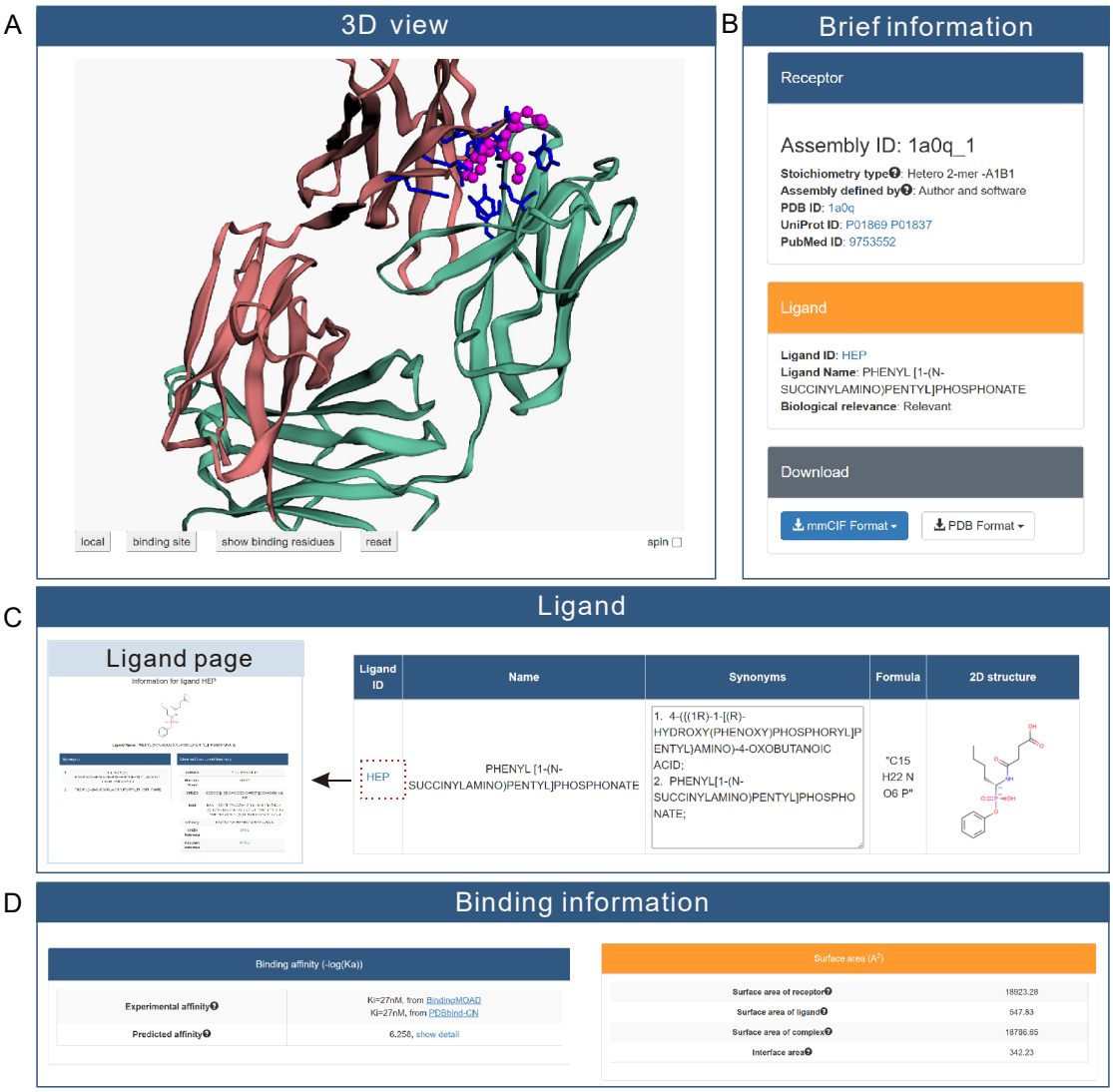

Figure 5. Information displayed on the webpage for an example BioLiP2 entry. (A) The 3D visualization of the ligand-protein interaction. (B) The overall brief information about Receptor, Ligand, and Download. (C) The detailed ligand information, including Ligand ID, Ligand Name, Synonyms, Formula, and 2D structure. (D) The binding affinity and surface area.