Nucleic Acids Research, 2024, **52**, D404–D412
https://doi.org/10.1093/nar/gkad630
Advance access publication date: 31 July 2023
Database issue

BioLiP2: an updated structure database for biologically relevant ligand–protein interactions

Chengxin Zhang [1], Xi Zhang [2], Lydia Freddolino [1,2,*] and Yang Zhang [1,2,3,4,5,*]

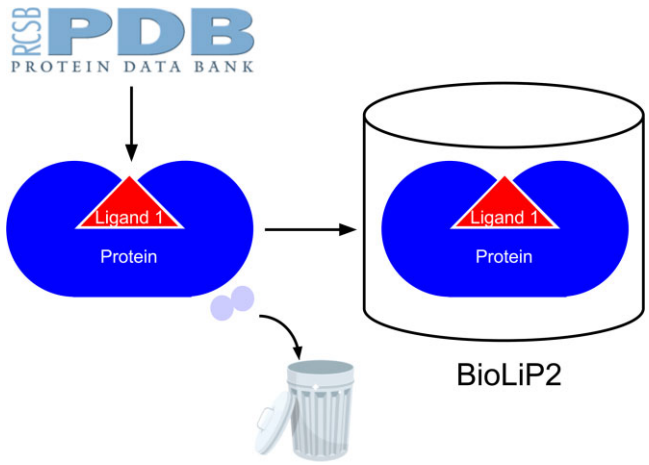
[1]Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI 48109, USA
[2]Department of Biological Chemistry, University of Michigan, Ann Arbor, MI 48109, USA
[3]Department of Computer Science, School of Computing, National University of Singapore, 117417, Singapore
[4]Cancer Science Institute of Singapore, National University of Singapore,117599, Singapore
[5]Department of Biochemistry, Yong Loo Lin School of Medicine, National University of Singapore, 117596, Singapore

*To whom correspondence should be addressed. Tel: +1 734 647 5839; Fax: +1 734 615 6553; Email: zhang@zhanggroup.org
Correspondence may also be addressed to Lydia Freddolino. Email: lydsf@umich.edu

Abstract

With the progress of structural biology, the Protein Data Bank (PDB) has witnessed rapid accumulation of experimentally solved protein structures. Since many structures are determined with purification and crystallization additives that are unrelated to a protein's *in vivo* function, it is nontrivial to identify the subset of protein–ligand interactions that are biologically relevant. We developed the BioLiP2 database (https://zhanggroup.org/BioLiP) to extract biologically relevant protein–ligand interactions from the PDB database. BioLiP2 assesses the functional relevance of the ligands by geometric rules and experimental literature validations. The ligand binding information is further enriched with other function annotations, including Enzyme Commission numbers, Gene Ontology terms, catalytic sites, and binding affinities collected from other databases and a manual literature survey. Compared to its predecessor BioLiP, BioLiP2 offers significantly greater coverage of nucleic acid-protein interactions, and interactions involving large complexes that are unavailable in PDB format. BioLiP2 also integrates cutting-edge structural alignment algorithms with state-of-the-art structure prediction techniques, which for the first time enables composite protein structure and sequence-based searching and significantly enhances the usefulness of the database in structure-based function annotations. With these new developments, BioLiP2 will continue to be an important and comprehensive database for docking, virtual screening, and structure-based protein function analyses.

Graphical abstract



Introduction

Although the Protein Data Bank (PDB) (1) provides rich structural information for proteins, it hosts limited functional annotations. For example, it is difficult to identify which ligands are biologically relevant to protein functions, versus which are simply additives used solely for protein purification and crystallization purposes. This has made it hard to curate the protein–ligand interaction information in the PDB for protein–ligand docking and template-based function annotations. Moreover, the PDB includes almost no protein-level function annotations such as Gene Ontology (GO) terms (2), making it even harder to understand the protein–ligand interactions in the context of molecular function or biological pathways.

Received: June 1, 2023. **Revised:** July 3, 2023. **Editorial Decision:** July 6, 2023. **Accepted:** July 17, 2023

To compensate for the lack of functional annotations in the PDB, many secondary databases have been developed, each with their own advantages and limitations. For example, the SIFTS (3) and PDBsum (4) databases annotate PDB proteins with GO terms (2) and Enzyme Commission (EC) numbers (5) by mapping them to functionally annotated sequences and domains in the UniProt (6) and InterPro (7) databases, respectively. However, both SIFTS and PDBsum do not contain ligand interactions. LigASite (8) and DUD-E (9) are two manually curated datasets of protein–ligand interactions, but they each contain only a few hundred proteins. The BindingDB (10), Binding MOAD (11) and PDBbind-CN (12) are three relatively comprehensive databases for binding affinities associated with protein–ligand interactions found in PDB structures, but they do not include protein–ligand interactions lacking known binding affinities. PepBDB (13) and RsiteDB (14) are structural databases of protein–peptide and protein–RNA interactions, and do not contain information on regular ligands. The DescribePROT database includes predicted protein–protein, protein–RNA and protein–DNA interactions, but does not include protein-small molecule interactions (15). MobiDB (16) includes predicted and experimental interaction information within intrinsically disordered regions of proteins, but the collected interactions are mainly for protein–protein interactions. The PDBe-KB database aggregates residue-level function annotations, including ligand binding sites and post-translational modification sites, for all proteins in the PDB (17). However, it does not differentiate between biologically relevant versus irrelevant ligand–protein interactions. FireDB (18) is a database for interaction between proteins and small-molecule ligands, and probably the only other database (apart from the BioLiP (19) database that we previously developed) that differentiate biologically relevant versus biologically irrelevant interactions. However, FireDB identifies biologically relevant interactions, which are referred to as the 'cognate' set, only by manual annotation of ligands through literature search, and hence cannot keep up with the constant growth of the PDB database. This leaves biological relevance of most ligand–protein interactions, which are critically important to properly annotating protein functions, unknown. Finally, IBIS (20) is another database for protein–ligand interaction, where the biological relevance of ligand binding site is assessed by a combination of geometry rules (atomic distances and number of contacts) and evolutionary conservation of binding residues. Unfortunately, IBIS has ceased update since 2017, making its dataset obsolete.

To address the critical gap noted above in high-quality functional annotations of ligand binding interactions in the PDB, we previously developed BioLiP (19), which is a comprehensive structure database for biologically relevant ligand–protein interactions. BioLiP assesses the biological relevance of ligand–protein interactions by a composite pipeline consisting of geometric rules, common addictive filters, and validation from experimental literature. The ligand binding annotations are complemented by other data sources to provide comprehensive protein function annotations, including ligand binding affinities collected from BindingDB, MOAD and PDBbind-CN databases as well as manual survey of literature; GO and EC annotations from the SIFTS database; and catalytic residues annotated by the M-CSA (previously CSA) database (21) and cross-links to other databases. These detailed functional annotations have made BioLiP useful to many structure-function studies, including binding site prediction (22–25), virtual screening (26–28), docking (29–31), and protein function predictions (32–35).

Here, we report a newly extended database, BioLiP2, which is developed to include seven major improvements to address issues with BioLiP and to significantly enhance its utility for a broader community of biological users. First, a convenient and fast searching engine is critical to any databases. While BioLiP only contains search through protein and ligand IDs which limit the search capacity, BioLiP2 extends the search to both protein/ligand sequence and structure search by the integration of the quick structure alignment algorithms (36,37) and a new AI-based protein structure library (38). Second, for protein-nucleic acid interactions, rather than only collecting short DNAs and RNAs with <30 nucleotides as in BioLiP, BioLiP2 includes all DNA–protein and RNA–protein interactions present in the PDB structures covered. Third, whereas BioLiP does not contain sequence information for peptide and nucleic acid ligands, BioLiP2 displays the full sequences of these biopolymer ligands. Fourth, while BioLiP was based on protein–ligand interactions extracted from legacy PDB format coordinate files, BioLiP2 extracts interactions from the macro-molecular Crystallographic Information File (mmCIF) format coordinate files from the PDB database, enabling it to parse >3000 structures without PDB format files in the PDB database. For example, BioLiP2 contains 3385 protein–ligand interactions from 812 protein chains extracted from mmCIF file of the phycobilisome structure (PDB 5y6p, https://zhanggroup.org/BioLiP/qsearch.cgi?&page=last&order=pdbid&pdbid=5y6p), which was too large to be recorded by the PDB format in the PDB database and thus not covered by the old BioLiP database. Fifth, in addition to simply listing GO terms annotated to a protein partner as in BioLiP, BioLiP2 additionally draws the directed acyclic graphs (DAGs) for all GO terms and their parent terms for the protein, allowing for easier visualization of the relations among different proteins. Sixth, more detailed small molecule information is displayed, including the IUPAC International Chemical Identifier (InChI) (39), InChIKey, SMILES string, and crosslinks to external small molecule databases, including ChEMBL(40), DrugBank (41) and ZINC (42). Last but not the least, all underlying source code for database curation and web interface display are made available under the BSD open-source license. To our knowledge, BioLiP2 represents the largest database of biologically relevant protein–ligand interactions, and the only such database that makes its curation code open-source.

Materials and methods

BioLiP2 extracts biologically relevant protein–ligand interactions from experimental structures (Figure 1). In the first step, the mmCIF files of all protein-containing structures are downloaded from the PDB database and split into chains by a modified version of the BeEM tool (43). Each chain is further split into a macromolecule part and small molecule part, where the former and the latter have numerical values and '.', respectively, in the 'label_seq_id' record of the mmCIF file. Non-standard residues from the macromolecules are then mapped to standard residue types (https://zhanggroup.org/BioLiP/help.html#TextS1).

In the second step, all ligands from all chains in the same mmCIF file are collected, including small molecules (including metal ions), peptides with < 30 amino acids, DNAs, and

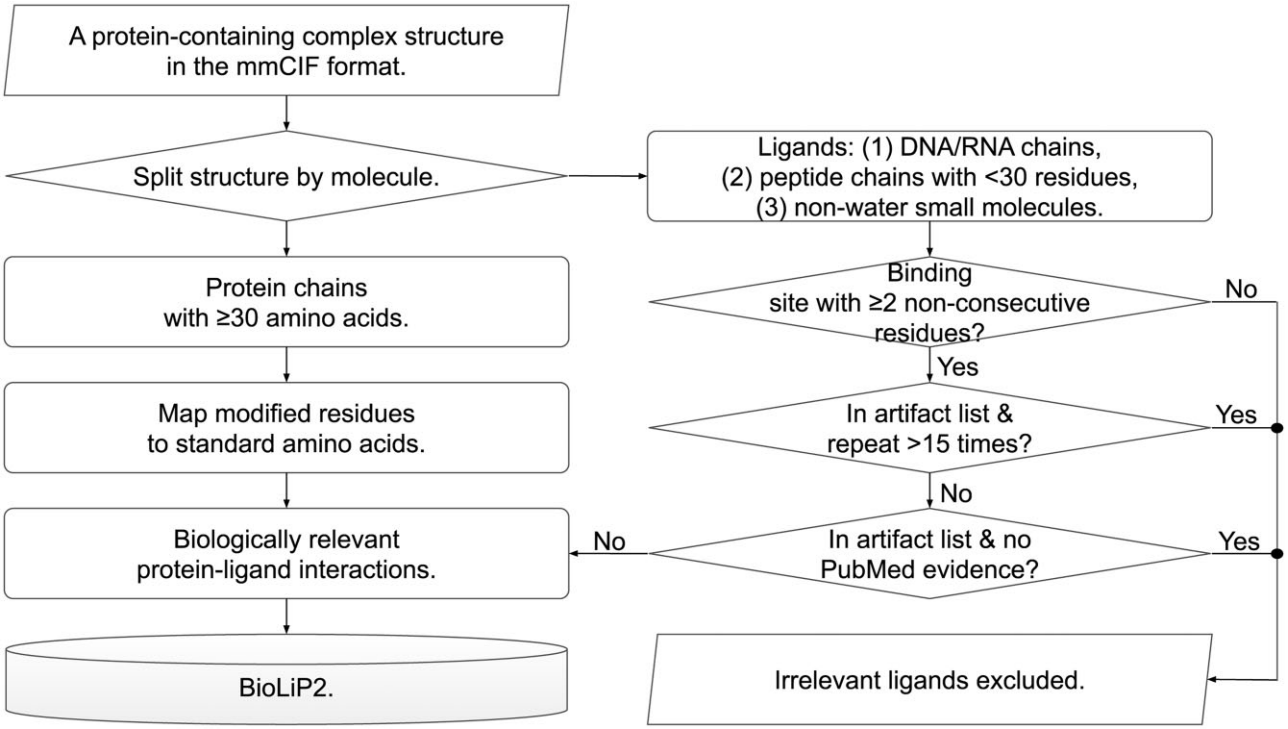


Figure 1. Flowchart for identification of biologically relevant ligand–protein interactions in BioLiP2.

RNAs. All protein components with ≥30 amino acids are also collected (differentiated from peptides, which are treated as ligands, by their length). For each protein–ligand pair, all inter-molecular atomic contacts, i.e. protein–ligand atom pairs within sum of van der Waals radii plus 0.5 Å, are calculated among non-hydrogen atoms. A protein residue with at least two inter-molecular atomic contacts to a ligand is labeled as a ligand binding residue. Any group of two or more ligand binding residues for the same protein–ligand pair are grouped into a binding site. Ligands without a binding site on any protein chains are excluded.

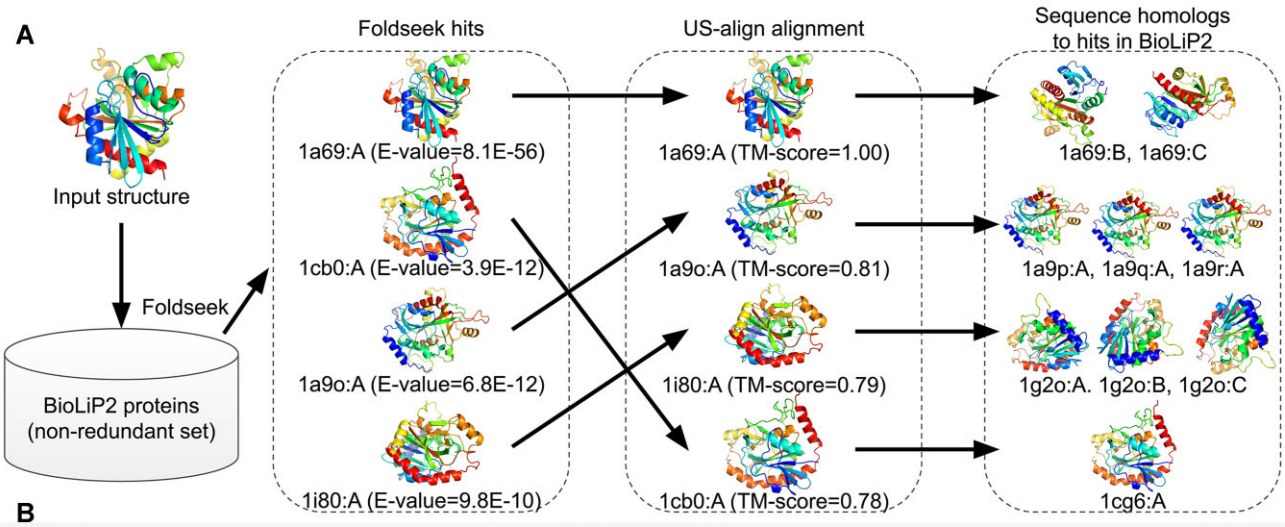
In the third step, biological relevance of each ligand is further assessed if it is one of the 463 commonly used non-biological ligands (https://zhanggroup.org/BioLiP/ligand_list). This list of potential artifact ligands includes compounds frequently used for crystallization additives and protein purification buffers. Among them, 353 ligands were collected by our previous study (19) describing the first version the BioLiP database and the remaining 110 were added over the years. A ligand on the artifact list will be discarded if it appears >15 times in the structure, or if its binding site only contains two consecutive residues. Otherwise, the title and abstract of the primary citation describing the structure is downloaded from PubMed to check the remaining potential artifacts. Here, the PubMed ID is extracted from the 'pdbx_database_id_PubMed' field of the mmCIF file when available; otherwise, the mapping between PDB and PubMed provided by the SIFTS database (3) is used. If the chemical name or synonyms of the ligand are found in the PubMed abstract or title, it is deemed biologically relevant; otherwise, it is discarded as irrelevant.

Finally, for each protein chain with at least one biologically relevant ligand, its mapping to UniProt proteins, GO terms, EC numbers, and species of origin are extracted from the mapping files provided by the SIFTS database (3). Since SIFTS only provides leaf GO terms without their parent GO terms in the GO annotations, in-house code is used to derive all parent GO terms for every leaf GO terms annotated to a protein chain so that users can search either the leaf terms or the parent terms through the database. The protein names and gene names for the proteins are extracted from their corresponding UniProt entries. The catalytic site residues are provided by the M-CSA (21) database. Binding affinities for each ligand are collected from BindingDB (10), Binding MOAD (11) and PDBbind-CN (12) databases. Additionally, a small number (81 cases) of binding affinities are collected from a manual survey of experimental literature performed in our previous study (19). The name, synonyms, chemical formula, and linear descriptions of a small molecule ligand, including SMILES, InChI and InChIKey, are extracted from the Chemical Component Dictionary (CCD) provided by the PDB database. Mappings from PDB ligand IDs to ligand IDs in ChEMBL, DrugBank and ZINC databases are performed using the UniChem database (44).

Results

BioLiP2 in numbers

BioLiP2 is updated weekly, usually on Friday, to keep up with the release cycle of the PDB database. At the time of writing of this article, BioLiP2 contains 385 160 protein chains involved in 781 684 protein–ligand interactions, including 35 167 (4%), 36 784 (5%), 127 525 (33%), 174 257 (45%) and 40 7951 (52%) interactions with peptides, DNAs, RNAs, metal ions, and other small molecules (which are referred to as 'regular' ligands by BioLiP2), respectively. Among all these interactions, 50064 have binding affinity data from, including



B

Search input.pdb through BioLiP using Foldseek and US-align (-fast mode). Results are ranked in descending order of TM-score normalized by query.

#	Hit (length)	Aligned length	TM-score (Identity) normalized by query	TM-score (Identity) normalized by hit	RMSD (Identity) normalized by aligned length	Foldseek E-value	Homologs to hit
1	1a69:A (237)	237	1.0000 (1.000)	1.0000 (1.000)	0.00 (1.000)	8.138E-56	1a69:C, 1a69:B
2	1a9o:A (289)	233	0.8080 (0.118)	0.6780 (0.097)	3.34 (0.120)	3.900E-12	1a9p:A, 1a9q:A, 1a9r:A, 1a9s:A, 1a9t:A, 1b8n:A, 1b8o:A, 1fxu:A
3	1i80:A (266)	227	0.7873 (0.089)	0.7115 (0.079)	3.31 (0.093)	9.830E-10	1g2o:A, 1g2o:B, 1g2o:C, 1i80:B, 1i80:C
4	1cb0:A (268)	220	0.7754 (0.127)	0.6951 (0.112)	3.15 (0.136)	6.779E-12	1cg6:A

Figure 2. Browsing interactions for regular ligands (**A**) and RNA ligands (**B**) in BioLiP2. More details of the function annotations are displayed by hovering over different fields. In this example, hovering over the 'GO terms' field results in a full list of GO terms and names associated with the protein receptor, where F, P and C stand for Molecular Function, Biological Process and Cellular Component (aspects of GO terms), respectively.

26 825, 25 693, 19 922 and 81 from MOAD, PDBbind-CN, BindingDB and direct literature search, respectively. Compared to the last release of BioLiP in 2022 with 573 225 entries, BioLiP2 increases the database size by 36% due to including large nucleic acids, extracting interactions from mmCIF rather than PDB format structure files, and more up-to-date data.

BioLiP2 web interface

The BioLiP2 web interface at https://zhanggroup.org/BioLiP provides four basic interfaces: SEARCH, BROWSE, LIGAND and DOWNLOAD. The following sections describe each interface in detail.

Searching BioLiP2

The 'SEARCH' interface provides three approaches to search the BioLiP2 database: 'search by name', 'search by sequence' and 'search by structure'. First, the user can query BioLiP2 by PDB ID, PDB chain ID, ligand ID (the 3-letter code defined by the Chemical Compound Dictionary by the PDB database), ligand name, UniProt accession, EC number, GO term, or PubMed ID.

Second, BioLiP2 provides a new functionality to search entries by the protein or nucleic acid sequences of either the protein partner or biopolymer ligands (peptides or nucleic acids). For moderate to long query sequences with $\geq$ 30 residues, NCBI BLAST+ (45) is used to search a local non-redundant sequence database clustered at 90% (for proteins) or 100% (for RNAs, DNAs and peptides) sequence identity cutoff. For short queries with $<$ 30 residues, Needleman-Wunsch sequence alignment (46) is used as BLAST + often fail to return any hit for short queries even if there is an identical match. In the search results, both representative hits found in the non-redundant database and members in the same sequence clusters are listed.

Third, BioLiP2 provides offers another new functionality to search entries by protein structure. In this approach, the user provides the input structure by uploading the structure file in PDB (or mmCIF) format, by specifying the PDB ID and chain ID, or by providing the UniProt accession; in the last case, the input structure model will be downloaded from the AlphaFold DB (38). The input structure will be quickly scanned through the non-redundant set of BioLiP2 receptors using Foldseek (36). Significant (E-value $\leq$ 0.001) Foldseek hits will be realigned by US-align (37) to calculate the TM-score (47) between input and BioLiP2 structure (Figure 2).

In addition to the 'SEARCH' interface, BioLiP2 can also be queried by RESTful API as documented by https://zhanggroup.org/BioLiP/help.html#api. The API can return search result either in HTML or plain text format.

Browsing BioLiP2

BioLiP2 can be browsed by protein–ligand interactions and by ligands through the 'BROWSE' and 'LIGAND' interfaces, respectively. The 'BROWSE' interface displays the PDB ID and chain ID, resolution, ligand, EC number, GO terms, UniProt accessions, PubMed citations and binding affinities, either for all protein–ligand interactions or for the subset of interactions with regular ligands, metal ions, RNAs, DNAs, and peptides (Figure 3A). For RNAs, DNAs and peptides, the full sequences are displayed (Figure 3B). The RNA secondary structures assigned by CSSR (48) are also shown in dot bracket format (Figure 3B column 4). More information such as protein name, ligand binding residues and enzymatic activity descriptions can be viewed by hovering over each entry. Clicking on

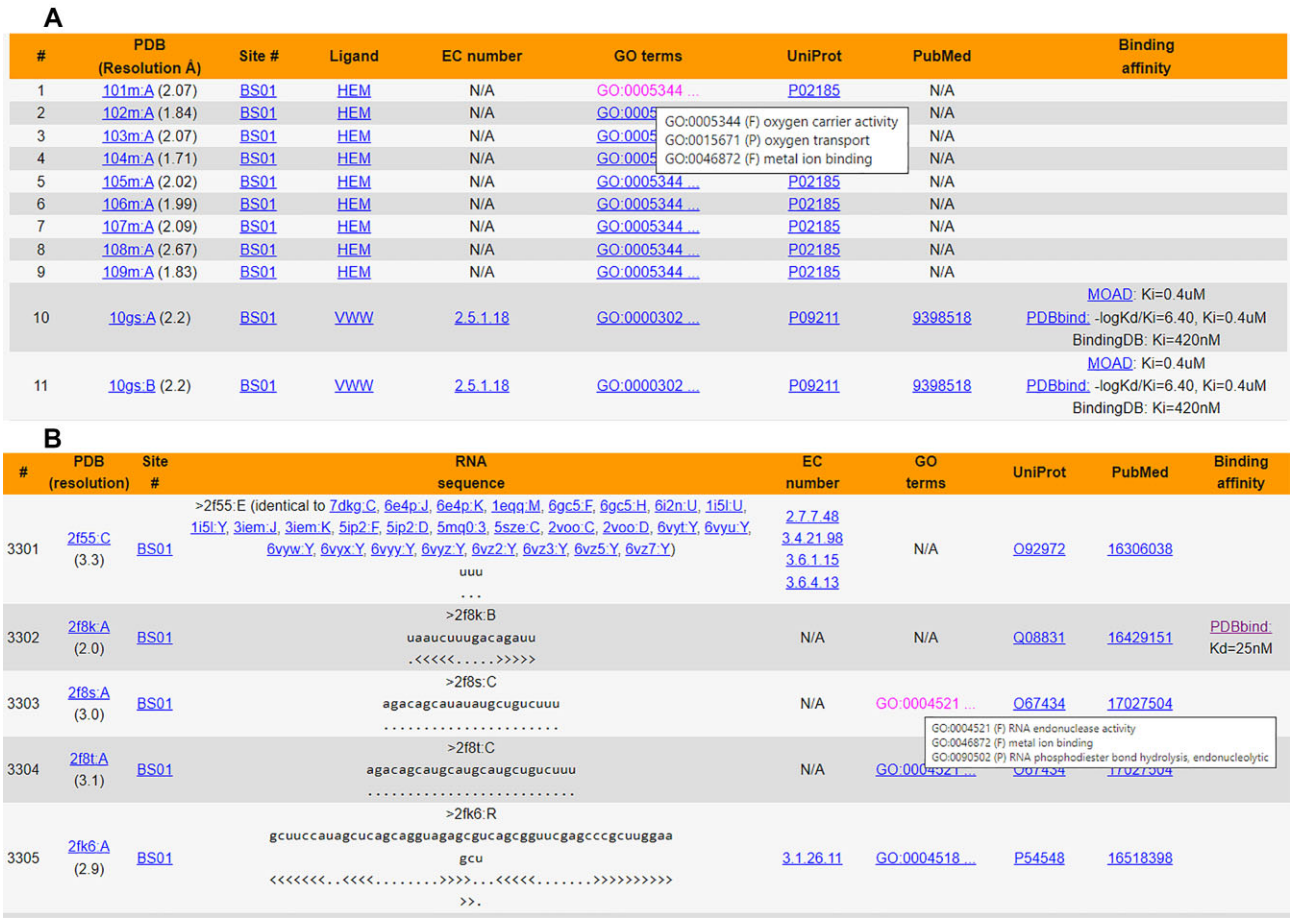


Figure 3. Browsing interactions for regular ligands (**A**) and RNA ligands (**B**) in BioLiP2. More details of the function annotations are displayed by hovering over different fields. In this example, hovering over the 'GO terms' field results in a full list of GO terms and names associated with the protein receptor, where F, P and C stand for Molecular Function, Biological Process and Cellular Component (aspects of GO terms), respectively.

the 'Site' column will lead to an individual page for detailed structure and function information for the interaction.

This page includes the sequence and structure of the protein–ligand complex and for the ligand binding site, which are displayed by two JSmol (49) applets (Figure 4). For enzymes, EC numbers and enzyme names as well as catalytic sites are displayed when available. GO terms for Molecular Function, Biological Process and Cellular Component are displayed in three separate tables, followed by three directed acyclic graphs drawn by Graphviz (50) to show the relations among different GO terms. Miscellaneous information such as resolution, name of the structure, protein name, species, and crosslinks to other databases such as PDB, PDBsum, PubMed and UniProt are also included. For a small molecule ligand, this page displays the 2D diagram, ligand IDs (PDB CCD ID, ChEMBL ID, DrugBank ID and ZINC ID), chemical formula, ligand name, and linear descriptions of the molecules (InChI, InChIKey, and one or more SMILES strings).

BioLiP2 can also be browsed by ligands through the 'LIGAND' interface, which displays the ligand IDs, chemical formula, ligand name, linear descriptions of the molecules, and a link to all BioLiP2 interactions associated with the ligand.

Downloading BioLiP2 data

The 'DOWNLOAD' interface of BioLiP2 allows batch download of all its data, including the protein and ligand chains, ligand IDs, ligand binding and catalytic residues, binding affinities, GO terms, EC numbers, PubMed citations and all structure files for receptor and ligand interactions. Two versions of the datasets are provided: a redundant dataset with all data, and a non-redundant dataset where receptor proteins share < 90% sequence identity as calculated by CD-HIT (38). The FASTA sequences of protein receptors as well as peptides, RNAs and DNAs are also available. Meanwhile, the source codes for database curation and web interface display are also available in this page.

Discussion and conclusion

We developed BioLiP2, a significantly extended version of the popularly used BioLiP protein function database. The extensions are focused on three aspects of improvements. First, BioLiP2 significantly expands the coverage of BioLiP by including more comprehensive DNA–protein and RNA–protein interactions, as well as interactions found only in mmCIF format entries but not PDB format entries in the PDB. Second, by integrating a cutting-edge protein structural alignment algorithm and state-of-the-art protein structure predictions, BioLiP2 offers the previously unattainable capacity for structure-based database search and enhances the search capacity of structure-based function annotation. Finally, BioLiP2 im-

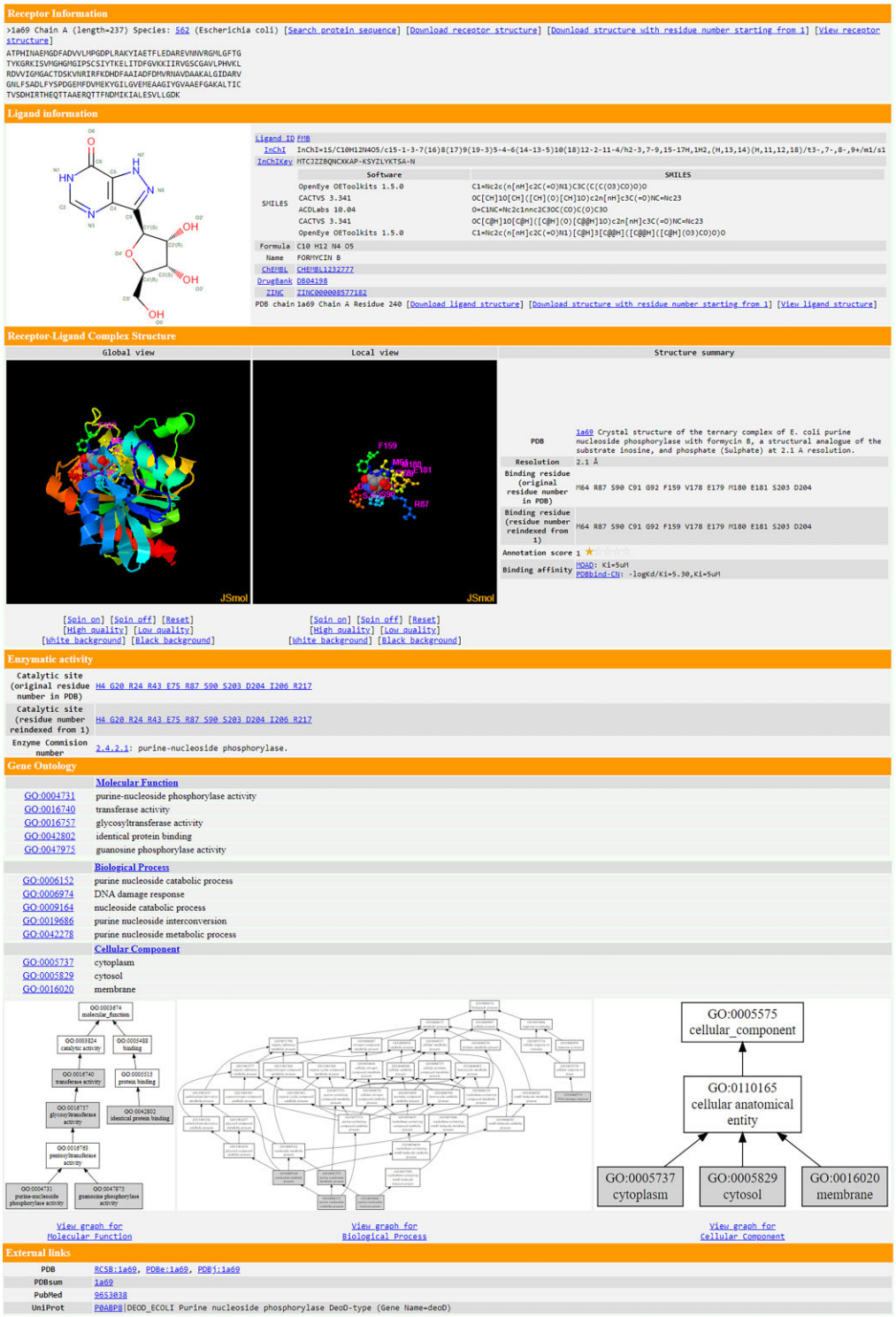


Figure 4. An example of detailed BioLiP2 annotations for the interaction between formycin B and purine nucleoside phosphorylase DeoD from *E. coli* (PDB ID: 1a69 chain A, https://zhanggroup.org/BioLiP/pdb.cgi?pdb=1a69&chain=A&bs=BS01). In addition to the receptor protein sequence and chemical information for the ligand, this page also shows the global and local structure of the ligand–protein complex using a JSmol applet. This is followed by the list of protein-level function annotations (EC numbers and GO terms) with directed acyclic graphs plotted for each of the three GO aspects. In the GO graphs, GO terms shaded in grey are GO terms directly annotated by the SIFTS database, while their parent terms are unshaded. The protein name, gene name, and crosslinks to external databases are listed at the end of the page. Each protein–ligand interaction is assigned an annotation score ranging from 1 to 5 shown at the end of the 'Receptor-Ligand Complex Structure' section. Higher annotation scores suggest greater biological relevance. If the UniProt protein for the receptor chain is mapped to at least one Rhea reaction, all non-water substrates and products of the reaction(s) are converted to 1024-bit Morgan fingerprints (ECFP4). Their chemical similarity to the ligand in question can then be measured by Tanimoto Coefficient (TC). The highest TC among all substrates/products to the ligand is used to assign the annotation score: TCs in the ranges [0,0.4), [0.4,0.6), [0.6,0.8), [0.8,1) and 1 correspond to annotation scores of 1, 2, 3, 4 and 5, respectively. If the receptor protein cannot be mapped to Rhea, the annotation score is assigned based on FireDB classification of ligands, where 'cognate', 'ambiguous' and 'non-cognate' ligands are assigned scores of 1, 3 and 4, respectively. In the above example, the ligand (formycin B) is not the native ligand (inosine) but its analog, the annotation score is low.

proves the effectiveness and user-friendliness of the web interface with including new features of Gene Ontology graphs and the source codes for database curation and web interface display.

With the new developments and enhancements of the database, which represent by far the largest library of biologically relevant protein–ligand interactions, BioLiP2 will continue to serve the broader biomedical community as an important database for protein–ligand docking, virtual ligand-screening, and structure-based protein function annotations. We expect that BioLiP2 will provide substantially improved utility both as a database enabling other tools, and for interactive use. Nevertheless, as the name suggests, a current limitation of BioLiP2 is that it is centered around protein–ligand interactions. Future versions of BioLiP2 will include RNA-small molecule interactions, given the increasingly recognized importance of the latter in drug discovery (51). Since BioLiP2 focuses on protein–ligand information, it does not collect protein–protein interactions except for interactions with short peptides. As a complementary resource, we are developing HomodimerDB (https://seq2fun.dcmb.med.umich.edu/HomodimerDB/) which will be a comprehensive and non-redundant database of homomeric protein–protein interactions.

During the peer review of this work, the authors become aware of a similar database of protein–ligand interactions separately developed by Wei et al. (https://yanglab.qd.sdu.edu.cn/Q-BioLiP/). While both databases were extensions of the original BioLiP database, the two resources are complementary to each other in terms of both web interface and underlying data. Our BioLiP2 database focuses on improving the usability of the database with newly added sequence/structure capabilities and providing a comprehensive set of biologically relevant ligand–protein pairs in the PDB database. Meanwhile, the database from Wei *et al.* focuses on protein–ligand interactions in the context of oligomeric protein complexes. Therefore, both databases provide differing and important tools to the biological community.

Data availability

The BioLiP2 database and source code are available at https://zhanggroup.org/BioLiP/ and https://github.com/kad-ecoli/mmCIF2BioLiP (permanent doi: https://doi.org/10.6084/m9.figshare.23641701) under the BSD license. Scripts for database curations are written in Perl 5 and C++ 11. Web interface are written in HTML and Python 3. All in-house Perl, C++ and Python code use standard libraries without external library dependencies, and thus should be compatible with any UNIX-like systems.

Supplementary data

Supplementary Data are available at NAR Online.

Acknowledgements

The authors thank Dr Jianyi Yang for technical assistance. The authors thank Dr Xiaoqiong Wei and Mr. Qingyuan Liu for testing BioLiP2. This work used the Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS) program, which is supported by National Science Foundation [2138259, 2138286, 2138307, 2137603, and 2138296]. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Funding

National Institute of General Medical Sciences [GM136422 and S10OD026825 to Y.Z.]; National Institute of Allergy and Infectious Diseases [AI134678 to L.F. and Y.Z.]; National Science Foundation [IIS1901191 and DBI2030790 to Y.Z. and MTM2025426 to L.F. and Y.Z.]. Funding for open access charge: National Institutes of Health.

Conflict of interest statement

None declared.

References

1. Berman,H., Henrick,K. and Nakamura,H. (2003) Announcing the worldwide Protein Data Bank. *Nat. Struct. Biol.*, **10**, 980.
2. Gene Ontology,C. (2021) The Gene Ontology resource: enriching a GOld mine. *Nucleic Acids Res.*, **49**, D325–D334.
3. Dana,J.M., Gutmanas,A., Tyagi,N., Qi,G., O'Donovan,C., Martin,M. and Velankar,S. (2019) SIFTS: updated Structure Integration with Function, Taxonomy and Sequences resource allows 40-fold increase in coverage of structure-based annotations for proteins. *Nucleic Acids Res.*, **47**, D482–D489.
4. Laskowski,R.A., Jablonska,J., Pravda,L., Varekova,R.S. and Thornton,J.M. (2018) PDBsum: structural summaries of PDB entries. *Protein Sci.*, **27**, 129–134.
5. Bairoch,A. (2000) The ENZYME database in 2000. *Nucleic Acids Res.*, **28**, 304–305.
6. UniProt,C. (2021) UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res.*, **49**, D480–D489.
7. Blum,M., Chang,H.Y., Chuguransky,S., Grego,T., Kandasaamy,S., Mitchell,A., Nuka,G., Paysan-Lafosse,T., Qureshi,M., Raj,S., *et al.* (2021) The InterPro protein families and domains database: 20 years on. *Nucleic Acids Res.*, **49**, D344–D354.
8. Dessailly,B.H., Lensink,M.F., Orengo,C.A. and Wodak,S.J. (2008) LigASite - a database of biologically relevant binding sites in proteins with known apo-structures. *Nucleic Acids Res.*, **36**, D667–D673.
9. Mysinger,M.M., Carchia,M., Irwin,J.J. and Shoichet,B.K. (2012) Directory of useful decoys, enhanced (DUD-E): better ligands and decoys for better benchmarking. *J. Med. Chem.*, **55**, 6582–6594.
10. Gilson,M.K., Liu,T., Baitaluk,M., Nicola,G., Hwang,L. and Chong,J. (2016) BindingDB in 2015: a public database for medicinal chemistry, computational chemistry and systems pharmacology. *Nucleic Acids Res.*, **44**, D1045–D1053.
11. Smith,R.D., Clark,J.J., Ahmed,A., Orban,Z.J., Dunbar,J.B. and Carlson,H.A. (2019) Updates to binding MOAD (Mother of all databases): polypharmacology tools and their utility in drug repurposing. *J. Mol. Biol.*, **431**, 2423–2433.
12. Liu,Z., Li,Y., Han,L., Li,J., Liu,J., Zhao,Z., Nie,W., Liu,Y. and Wang,R. (2015) PDB-wide collection of binding data: current status of the PDBbind database. *Bioinformatics*, **31**, 405–412.
13. Wen,Z., He,J., Tao,H. and Huang,S.Y. (2019) PepBDB: a comprehensive structural database of biological peptide-protein interactions. *Bioinformatics*, **35**, 175–177.

14. Shulman-Peleg,A., Nussinov,R. and Wolfson,H.J. (2009) RsiteDB: a database of protein binding pockets that interact with RNA nucleotide bases. *Nucleic Acids Res.*, **37**, D369–D373.
15. Zhao,B., Katuwawala,A., Oldfield,C.J., Dunker,A.K., Faraggi,E., Gsponer,J., Kloczkowski,A., Malhis,N., Mirdita,M., Obradovic,Z., *et al.* (2021) DescribePROT: database of amino acid-level protein structure and function predictions. *Nucleic Acids Res.*, **49**, D298–D308.
16. Piovesan,D., Necci,M., Escobedo,N., Monzon,A.M., Hatos,A., Micetic,I., Quaglia,F., Paladin,L., Ramasamy,P., Dosztanyi,Z., *et al.* (2021) MobiDB: intrinsically disordered proteins in 2021. *Nucleic Acids Res.*, **49**, D361–D367.
17. consortium,P.D.-K. (2022) PDBe-KB: collaboratively defining the biological context of structural data. *Nucleic Acids Res.*, **50**, D534–D542.
18. 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–D272.
19. Yang,J., Roy,A. and Zhang,Y. (2013) BioLiP: a semi-manually curated database for biologically relevant ligand–protein interactions. *Nucleic Acids Res.*, **41**, D1096–D1103.
20. Shoemaker,B.A., Zhang,D., Tyagi,M., Thangudu,R.R., Fong,J.H., Marchler-Bauer,A., Bryant,S.H., Madej,T. and Panchenko,A.R. (2012) IBIS (Inferred Biomolecular Interaction Server) reports, predicts and integrates multiple types of conserved interactions for proteins. *Nucleic Acids Res.*, **40**, D834–D840.
21. Ribeiro,A.J.M., Holliday,G.L., Furnham,N., Tyzack,J.D., Ferris,K. and Thornton,J.M. (2018) Mechanism and Catalytic Site Atlas (M-CSA): a database of enzyme reaction mechanisms and active sites. *Nucleic Acids Res.*, **46**, D618–D623.
22. Yang,J.Y., Roy,A. and Zhang,Y. (2013) Protein-ligand binding site recognition using complementary binding-specific substructure comparison and sequence profile alignment. *Bioinformatics*, **29**, 2588–2595.
23. Yu,D.J., Hu,J., Yang,J., Shen,H.B., Tang,J.H. and Yang,J.Y. (2013) Designing template-free predictor for targeting protein-ligand binding sites with classifier ensemble and spatial clustering. *IEEE ACM Trans. Comput. Biol. Bioinform.*, **10**, 994–1008.
24. Yuan,Q.M., Chen,S., Wang,Y., Zhao,H.Y. and Yang,Y.D. (2022) Alignment-free metal ion-binding site prediction from protein sequence through pretrained language model and multi-task learning. *Brief. Bioinform*, **23**, bbac444.
25. Santana,C.A., Izidoro,S.C., de Melo-Minardi,R.C., Tyzack,J.D., Ribeiro,A.J.M., Pires,D.E.V., Thornton,J.M. and de A. Silveira,S. (2022) GRaSP-web: a machine learning strategy to predict binding sites based on residue neighborhood graphs. *Nucleic Acids Res.*, **50**, W392–W397.
26. Roy,A., Srinivasan,B. and Skolnick,J. (2015) PoLi: a virtual screening pipeline based on template pocket and ligand similarity. *J. Chem. Inf. Model*, **55**, 1757–1770.
27. Litfin,T., Zhou,Y. and Yang,Y. (2017) SPOT-ligand 2: improving structure-based virtual screening by binding-homology search on an expanded structural template library. *Bioinformatics*, **33**, 1238–1240.
28. Zhang,W. and Huang,J. (2022) EViS: an enhanced virtual screening approach based on pocket-ligand similarity. *J. Chem. Inf. Model*, **62**, 498–510.
29. Wu,Q., Peng,Z.L., Zhang,Y. and Yang,J.Y. (2018) COACH-D: improved protein–ligand binding sites prediction with refined ligand-binding poses through molecular docking. *Nucleic Acids Res.*, **46**, W438–W442.
30. Zhang,W., Bell,E.W., Yin,M. and Zhang,Y. (2020) EDock: blind protein–ligand docking by replica-exchange monte carlo simulation. *J. Cheminform.*, **12**, 37.
31. 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.
32. Zhang,C., Freddolino,L. and Zhang,Y. (2017) COFACTOR: improved protein function prediction by combining structure, sequence and protein–protein interaction information. *Nucleic Acids Res.*, **45**, W291–W299.
33. Koo,D.C.E. and Bonneau,R. (2019) Towards region-specific propagation of protein functions. *Bioinformatics*, **35**, 1737–1744.
34. Smaili,F.Z., Tian,S., Roy,A., Alazmi,M., Arold,S.T., Mukherjee,S., Hefty,P.S., Chen,W. and Gao,X. (2021) QAUST: protein Function Prediction Using Structure Similarity, Protein Interaction, and Functional Motifs. *Genomics Proteomics Bioinformatics*, **19**, 998–1011.
35. Gligorijevic,V., Renfrew,P.D., Kosciolek,T., Leman,J.K., Berenberg,D., Vatanen,T., Chandler,C., Taylor,B.C., Fisk,I.M., Vlamakis,H., *et al.* (2021) Structure-based protein function prediction using graph convolutional networks. *Nat. Commun.*, **12**, 3168.
36. van Kempen,M., Kim,S.S., Tumescheit,C., Mirdita,M., Lee,J., Gilchrist,C.L.M., Söding,J. and Steinegger,M. (2023) Foldseek: fast and accurate protein structure search. *Nat. Biotechnol.*, https://doi.org/10.1038/s41587-023-01773-0.
37. Zhang,C., Shine,M., Pyle,A.M. and Zhang,Y. (2022) US-align: universal structure alignments of proteins, nucleic acids, and macromolecular complexes. *Nat. Methods*, **19**, 1109–1115.
38. Varadi,M., Anyango,S., Deshpande,M., Nair,S., Natassia,C., Yordanova,G., Yuan,D., Stroe,O., Wood,G., Laydon,A., *et al.* (2022) AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res.*, **50**, D439–D444.
39. Goodman,J.M., Pletnev,I., Thiessen,P., Bolton,E. and Heller,S.R. (2021) InChI version 1.06: now more than 99.99% reliable. *J. Cheminformatics*, **13**, 40.
40. Mendez,D., Gaulton,A., Bento,A.P., Chambers,J., De Veij,M., Felix,E., Magarinos,M.P., Mosquera,J.F., Mutowo,P., Nowotka,M., *et al.* (2019) ChEMBL: towards direct deposition of bioassay data. *Nucleic Acids Res.*, **47**, D930–D940.
41. Wishart,D.S., Feunang,Y.D., Guo,A.C., Lo,E.J., Marcu,A., Grant,J.R., Sajed,T., Johnson,D., Li,C., Sayeeda,Z., *et al.* (2018) DrugBank 5.0: a major update to the DrugBank database for. 2018. *Nucleic Acids Res.*, **46**, D1074–D1082.
42. Irwin,J.J., Tang,K.G., Young,J., Dandarchuluun,C., Wong,B.R., Khurelbaatar,M., Moroz,Y.S., Mayfield,J. and Sayle,R.A. (2020) ZINC20-A Free Ultralarge-Scale Chemical Database for Ligand Discovery. *J. Chem. Inf. Model*, **60**, 6065–6073.
43. Zhang,C. (2023) BeEM: fast and faithful conversion of mmCIF format structure files to PDB format. *BMC Bioinformatics*, **24**, 260.
44. Chambers,J., Davies,M., Gaulton,A., Hersey,A., Velankar,S., Petryszak,R., Hastings,J., Bellis,L., McGlinchey,S. and Overington,J.P. (2013) UniChem: a unified chemical structure cross-referencing and identifier tracking system. *J. Cheminform.*, **5**, 3.
45. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
46. Needleman,S.B. and Wunsch,C.D. (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.*, **48**, 443–453.
47. Zhang,Y. and Skolnick,J. (2004) Scoring function for automated assessment of protein structure template quality. *Proteins*, **57**, 702–710.
48. Zhang,C. and Pyle,A.M. (2022) CSSR: assignment of secondary structure to coarse-grained RNA tertiary structures. *Acta Crystallogr. D Struct. Biol.*, **78**, 466–471.
49. Hanson,R.M., Prilusky,J., Renjian,Z., Nakane,T. and Sussman,J.L. (2013) JSmol and the next-generation web-based representation of 3D molecular structure as applied to proteopedia. *Isr. J. Chem.*, **53**, 207–216.

50. Ellson,J., Gansner,E.R., Koutsofios,E., North,S.C. and Woodhull,G. (2004) Graphviz and dynagraph - static and dynamic graph drawing tools. In: Jünger,M. and Mutzel,P. (eds.) *Graph Drawing Software. Mathematics and Visualization*. Springer, Berlin, pp. 127–148.

51. Fedorova,O., Jagdmann,G.E., Adams,R.L., Yuan,L., Van Zandt,M.C. and Pyle,A.M. (2018) Small molecules that target group II introns are potent antifungal agents. *Nat. Chem. Biol.*, **14**, 1073–1078.

Received: June 1, 2023. **Revised:** July 3, 2023. **Editorial Decision:** July 6, 2023. **Accepted:** July 17, 2023