PDBBind Optimization to Create a High-Quality Protein-Ligand Binding Dataset for Binding Affinity Prediction

Yingze Wang*,[†] Kunyang Sun*,[†] Jie Li,[†] Xingyi Guan,[†] Oufan Zhang,[†] Dorian Bagni,[†] and Teresa Head-Gordon*,[†,‡]

[†]*Pitzer Center for Theoretical Chemistry and Department of Chemistry*
University of California, Berkeley, CA, USA 94720

[‡]*Departments of Bioengineering and Chemical and Biomolecular Engineering*
University of California, Berkeley, CA, USA 94720

E-mail: thg@berkeley.edu

Abstract

Development of scoring functions (SFs) used to predict protein-ligand binding energies requires high-quality 3D structures and binding assay data, and often relies on the PDBBind dataset for training and testing their parameters. In this work we show that PDBBind suffers from several common structural artifacts of both proteins and ligands and non-uniform reporting of binding energies of its derived training and tests, which may compromise the accuracy, reliability and generalizability of the resulting SFs. Therefore we have developed a series of algorithms organized in an automated workflow, PDBBind-Opt, that curates non-covalent protein-ligand datasets to fix common problems observed in the general, refined, and core sets of PDBBind. We also use PDBBind-Opt to create an independent data set by matching binding free energies

from BioLiP2 with co-crystalized ligand-protein complexes from the PDB. The resulting PDBBind-Opt workflow and BioLiP2-Opt dataset are designed to ensure reproducibility and to minimize human intervention, while also being open-source to foster transparency in the improvements made to this important resource for the biology and drug discovery communities.

*authors contributed equally

BACKGROUND & SUMMARY

Scoring functions (SFs) are crucial in computer aided drug discovery, utilized for selecting the most probable ligand geometry and its binding pose with a protein that best correlates or predicts their free energy of binding.[1] There are a plethora of SFs being developed and widely used by computational and medicinal chemists, and they can be broadly categorized into either classical scoring functions[2–12] or machine learning scoring functions.[13–20] The majority of protein-ligand SF predictors, whether physical or machine-learned, have been trained on the PDBBind dataset[21–27] (http://www.pdbbind-cn.org/), specifically v2020, a curated set of ~19,500 biomolecular complex structures and their experimentally measured binding affinities. PDBBind is further organized into a "general" data subset that is often adopted by SFs for training, and separate "refined" and "core" datasets which contain protein-ligand complexes with the best structural quality and most reliable binding affinity data that is used for testing. The Comparative Assessment of Scoring Functions (CASF) benchmark at 2016, which assesses the scoring power, ranking power, docking power and screening power of various SFs, was conducted on the PDBBind core set.[28,29]

PDBBind has been an invaluable resource to the biomolecular community and has been sustained for close to two decades. However, a significant portion of the PDBBind dataset has structural errors, statistical anomalies, and a sub-optimal organization of protein-ligand classes that can limit SF training and validation. Some of the structural issues are inherited from the the original RCSB Protein Data Bank (PDB)[30] dataset like missing hydrogen atoms

and/or incomplete residues due to uncertainties in the modeled electron densities, whereas some errors originate from the preparation of ligand structures that results in incorrect bond order, protonation state, tautomer state and aromaticity specifications. Some entries are covalent binders which require special methods to account for the covalent bond breaking and formulation,[31,32] and should remain distinct from protein-ligand complexes that are formed from non-covalent interactions only (Figure 1a). Upon close inspection, it has been found that not all structures in the refined set adhere to the strict criteria defined for their inclusion. For instance 5OUH[33] in the refined set is a noncovalent binder that exhibits a severe atomic clash with the protein (Figure 1b). Similarly, in 3KMC,[34] the chlorobenzene portion of the ligand is absent from the crystal structure (Figure 1c). These inconsistencies undermines the purpose of the refined set, which is intended to serve as a high-quality benchmark for evaluation of scoring functions and docking methods.
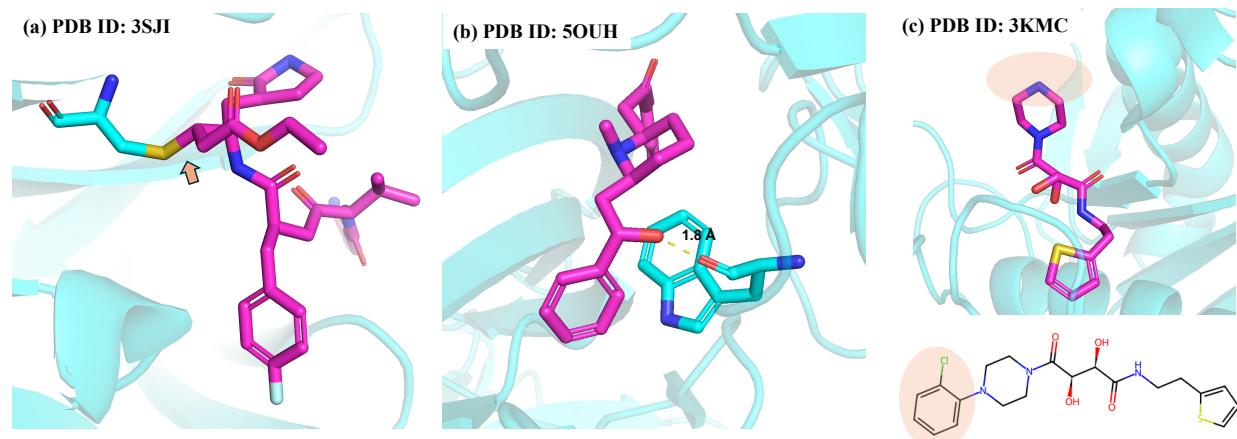

Figure 1: *Common structural imperfections in PDBBind dataset.* (a) Covalent binders. The ligand is covalently bound to cysteine with a Michael addition reaction. (b) Steric clashes with the distance between the clashing atoms being only 1.8 Å. (c) Missing atoms.

Another concern in regards PDBBind is that the data processing procedure is neither open-sourced nor fully automated, potentially relying on individual groups needing to introduce their own manual intervention that may lead to inconsistencies. Furthermore, the labor-intensive nature of the PDBBind data curation process became more problematic in

2021 when PDBBind ceased to be freely available, which limits access and hinders the development and validation of new scoring functions (and other additional uses). Therefore, there is a pressing need for an automated, open-sourced, reproducible, and systematic workflow to prepare high-quality protein-ligand binding structural datasets in order to foster greater transparency and accessibility.

In this work we introduce PDBBind-Opt, a workflow of automated algorithms for data cleaning and structural preparation that create a curated dataset of high-quality non-covalent protein-ligand complex structures with binding affinity annotations. This workflow incorporates several modules: (1) a curating procedure to reject any covalent bound complexes, ligands with very low frequency occurrences of certain atomic elements, and structures containing severe steric clashes; (2) a ligand fixing module to ensure the correctness of the ligand structure including correct bond order and reasonable protonation states; (3) a protein fixing module to extract all chains involved in the protein-ligand binding and to add missing atoms when necessary; (4) a structure refinement module to include hydrogens simultaneously within the protein-ligand complex state, as opposed to the current practice in PDBBind that completes the hydrogen chemistry for protein and ligand independently of each other. The motivation for adding this hydrogen growth module is that although many SFs only take heavy atoms into consideration, physics-based SFs may benefit from explicit hydrogens to better model intermolecular interactions such as hydrogen bonding.

To illustrate the utility of the PDBBind-Opt workflow, we create a new dataset of non-covalent small molecule ligand-protein complexes. Using the BioLiP2[35,36] resource that offers an expanded subset of ~48,000 non-covalent protein-ligand complexes, we extracted all the PDB IDs in this more complete collection that have binding data annotation and are not in the PDBBind-v2020 data set. Using our PDBBind-Opt workflow, we provide the structural annotations as described above to create the enhanced BioLiP2-Opt dataset that provides a new independent benchmark for physics-based and machine learning scoring functions. The PDBBind-Opt workflow and BioLiP2-Opt dataset are provided open-source to foster

4

transparency and sustainability as new data appears, in order to maintain this important resource for the biology and drug discovery communities.

METHODS

The flowchart of the PDBBind-Opt workflow is illustrated in Figure 2. We start by downloading the pdb and mmcif formats directly from the RCSB PDB[30] for each entry in PDBBind v2020. The pdb files are used for structure preparation and the headers in mmcif files are used to extract useful metadata, such as resolution, deposit date and sequence information. For each PDB entry, we split the structure into three components: ligand, protein and additives.

The following criteria is used to define the ligand(s): (1) Any residue will be identified as a ligand if its name matches the Chemical Component Dictionary (CCD) code deposited in PDBBind. Ligands identified in this manner are referred to as "small molecules". (2) If the ligand names in PDBBind contain patterns such as "*-mer," or symbols like "-", "&" or "+", we will select chains in the original PDB file that are less than 20 residues but more than one residue as ligands. These ligands are typically polypeptides, oligosaccharides, or oligonucleotides, collectively referred to as "polymers". For each identified ligand, we label any biopolymer chains within 10Å as the associated protein structure. Then, for each protein structure, we labeled residues specified by the "HETATM" record in the pdb file within 4Å as additives, which includes ions, solvents, and co-factors. The additives are saved in pdb format and directly deposited in the database, and the protein and ligand structure are ready to proceed to the next workflow steps.

After the structure splitting, a set of filters, including some borrowed from the processing protocols of LP-PDBBind,[37] are applied to exclude any protein-ligand complex structures that meet any of the following criteria:
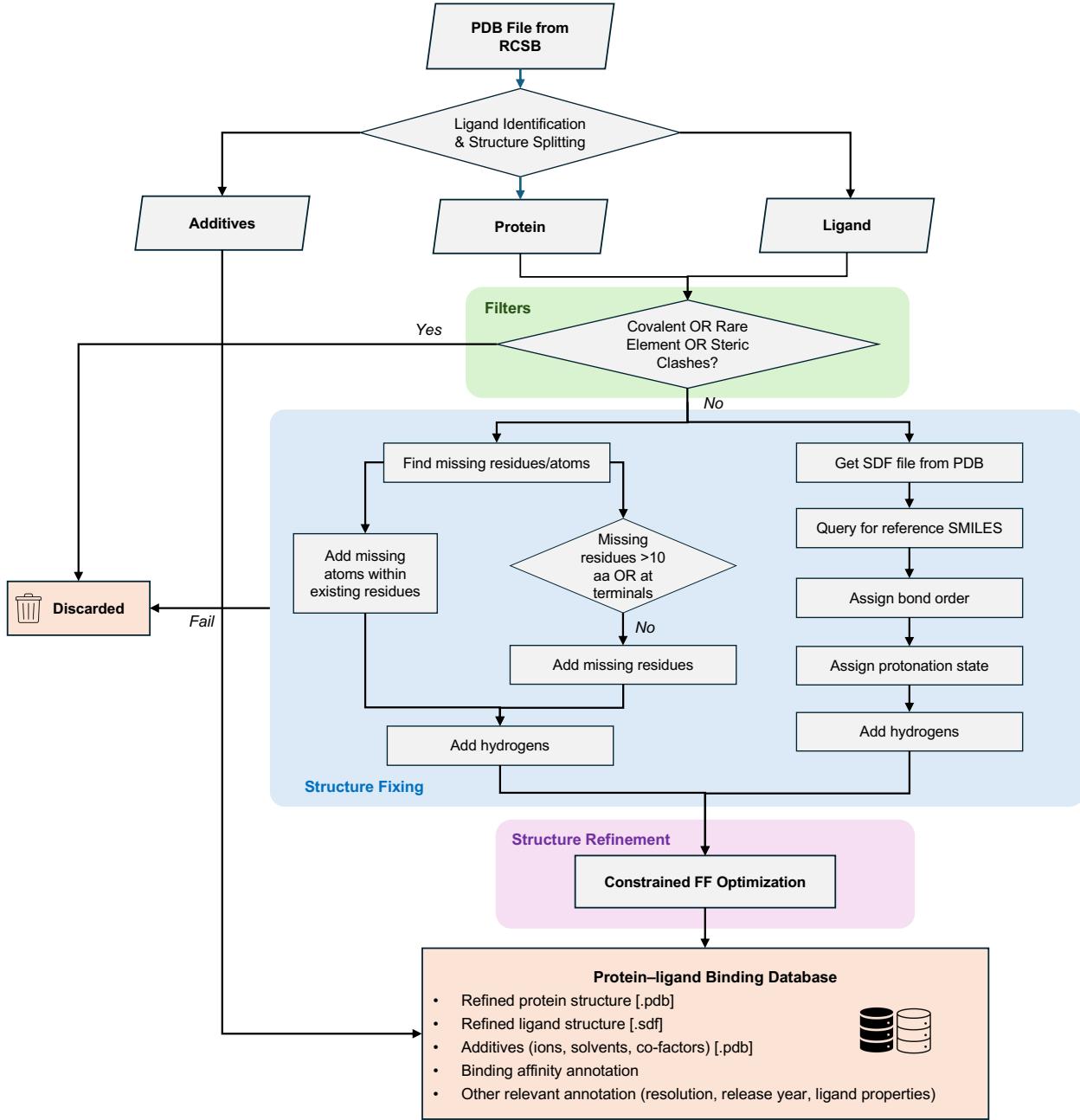
Figure 2: *Flowchart for PDBBind-Opt workflow.*

- Covalent binder filter: Excludes ligands covalently bound to the protein, as indicated by the "CONECT" record in the pdb file. When covalently-bound ligands are identified, they are eliminated (955 entries). The filter is designed because covalent binding inherently is different from non-covalent binding which does not involve bond breaking,

and thus requires special treatment in any SFs.

- <u>Rare element filter</u>: Excludes ligand containing elements other than H, C, N, O, F, P, S, Cl, Br, I. For example, Te or Se are infrequently encountered, and their inclusion can make it challenging for SFs to learn key binding features. This eliminated 205 entries from the original set of ligands.

- <u>Steric clashes filter</u>: Excludes structures with protein-ligand heavy atom pairs closer than 2 angstroms. Such steric clashes often arise from electron density uncertainties or inaccurate structural reconstruction from electron densities and are not physically feasible non-covalent interactions. Including such structures in SF development could be detrimental, for example leading to an underestimation of the repulsion energy in physics-based SFs. Additionally, the steric clash filter helps to exclude covalent ligands if the covalent bond is not properly represented in the "CONECT" record. This criteria filtered out 164 entries.

For protein-ligand complexes that pass these filters, two structure-fixing modules are implemented separately for proteins and ligands. In our ProteinFixer module, we first use the sequence information from the mmcif file header to detect missing atoms and residues. Then, for missing atoms within an existing residue or missing residues, PDBFixer[38] is used to add them, except when the missing residues are longer than 10 amino acids or are located at the sequence terminals. The final step of the protein fixing module is to add hydrogen atoms at pH=7.4 with PDBFixer. At this pH, all lysine (LYS) and arginine (ARG) are positively charged and glutamic acid (GLU) and aspartic acid (ASP) are deprotonated. Histidine remains in a neutral form and the HID or HIE variant is selected based on which one forms a better hydrogen bond.

Adding missing atoms to protein structure is essential because incomplete structures can compromise accurate modeling of binding interactions, and any molecular dynamics or alchemical binding free energy calculation also require complete structures to ensure the cor-

rect structural ensemble are sampled during simulations. Long missing segments or missing terminus residues in crystal structures, however, are often more variable such as intrinsically disordered region (IDR),[39] domains that are not expressed in the samples for crystallization, or his-tags introduced in the protein purification process.[40] But since these residues are often not directly involved in ligand binding, it should be safe to skip modeling these residues explicitly because otherwise it will introduce too much interference in the decisions about protein structure. Hence we leave these regions in their original form and in themselves do not define a criterion for being discarded in the final dataset.

In the LigandFixer module, we first obtain an sdf file for each ligand instance either by downloading from the RCSB PDB (if possible) or converting from the native pdb format with OpenBabel.[41] Since explicit atom connections are not present in the pdb format, the bond orders in this converted sdf file are typically inferred from local atomic geometries and the resulting structure is herein referred to as "inferred structure". Then, a reference SMILES is obtained, which is used to correct bond orders and aromaticity specifications that could sometimes be mislabeled in the inferred structure. The bond order assignment protocol is implemented as follows: if the inferred and reference structure are isomorphic, a one-to-one atom mapping will be generated by structure matching and then bond orders, atom hybridization and aromatic specifications will be assigned according to the reference; otherwise, the bond order assignment will come to a failure, which means that the inferred structure does not share the same number of atoms or bond connectivity as the reference, indicating that there are missing atoms or distorted geometries in the crystal structure. Therefore, such structures will be excluded.

One thing should be noted here is the source of the reference SMILES string. If the ligand is a small molecule with a CCD code or is a polymer with a BIRD (Biologically Interesting Molecule Reference Dictionary) code, we will query RCSB PDB for its reference SMILES. If the ligand is a polymer consisting only of alpha-amino acids, we will assume it is a simple non-cyclic peptide and generate a SMILES string based on its sequence information

8

and amide-bond formation rules. Apparently, for the latter case, any mismatch between the inferred and reference structure does not mean the inferred structure is wrong - the ligand may just be a cyclic peptide or contain disulfide bond. However, such structures will also be excluded as we are unable to verify its correctness automatically at this stage. For such cases, human inspection will be inevitable and its beyond the scope of the workflow.

After a correct structure is obtained, protonation states will be assigned and hydrogens are added according to a set of rules listed. We acknowledge that it is a non-trivial task to correctly determine protonation states for titratable groups within a ligand at a given pH and many algorithms that use empirical rules, QM/MM calculations or machine learning have been reported.[42–44] However, since our workflow is designed for high-throughput processing, we improve the efficiency using a simple set of predefined rules to determine the protonation states by relevant matching functional groups in SMARTS patterns. Acids, nitro groups, thiophenols, azides, and N-oxides are deprotonated. Aliphatic amines and guanidines/imines are protonated, while anilines are not protonated. There are other special considerations that should also be accounted for: Amines will not be protonated if the nitrogen is directly bonded with atoms other than H or C. Only one nitrogen atom on diamines and piperaizines will be protonated to avoid two positive charged groups close to each other, which is not favorable at normal biologically-relevant pH. Enols with the motif O=C-C=C-OH are deprotonated. The protonation state assignment is implemented by modifying the default behavior of `dimophite_dl` package[45] which can be found in the Github repository.

In addition we found that some of the SMILES strings deposited in RCSB PDB are incorrect such that all the bonds are labeled as single bonds. Most of these errors were caught by a geometric check for $sp^3/sp^2/sp$ carbons. For these cases, we manually corrected the SMILES according to the original literature and use the corrected one to do the ligand fixing. The list of manually corrected SMILES can be found in the public Github repository. The bond order assignment, protonation states assignement and hydrogen adding in the

ligand fixing module are all performed with RDKit.[46]

The last part of the PDBBind-Opt structural workflow is a refinement module in which the fixed ligand structure and protein structure are combined, followed by a constrained energy minimization with a well-established force field. AMBER14SB[47] is used for the protein and OpenFF-2.1.0[7] together with Gasteiger charges are used for ligand. The coordinate constraints are applied to all atoms that are experimentally resolved, which means only positions of hydrogens (both on the ligand and protein) and atoms added by PDBFixer in the protein fixing module are allowed to be optimized. We found this physically-based structural optimization is useful to resolve any remaining steric clashes between added atoms introduced by treating the protein and ligand structure separately in the previous structure fixing modules. Besides, hydrogen bonding network between the ligand and protein may also be optimized in this process. The constrained energy minimization was performed with OpenMM 8.1.1[48] by setting masses of all constrained atoms to zero. These final fixed pairs of protein and ligand structures define the PDBBind-Opt dataset.

The binding affinity in terms of ΔG is directly related to the dissociation coefficient K_d or K_i through the standard relationship $\Delta G = -RT \ln(K)$.[49] However, a large portion of the data in PDBBind is reported in terms of IC_{50}, which cannot be easily translated to ΔGs due to its dependence on other experimental conditions and inhibition mechanisms.[50] The IC_{50} values for the same protein-ligand complex can vary up to one order of magnitude in different assays, and some binding data in PDBBind are not reported as exact values but just ranges. Therefore, the binding affinity data is reorganized into a machine-readable format (csv) with comments as to the form of the experimental binding free energy data: K_d, K_i, and IC_{50}. Entries with EC_{50} are excluded because EC_{50} is usually measured in a cellular assay, and this value is affected by many factors other than protein-ligand binding affinity, such as a compound's stability and permeability.

TECHNICAL VALIDATION

Numbers of PDBBind-Opt dataset. Among the 19,443 unique PDB entries for proteins with non-covalent ligands in PDBBind v2020,[27] 15,589 of them were successfully processed using the PDBBind-Opt workflow (Figure 2); 1,324 entries were discarded by the filters and 2,580 structures were discarded because they were unable to pass the structure fixing and refinement modules. A large portion of the discarded datapoints are polymers in which it is are hard to verify its structural correctness. Almost certainly human inspection and expertise will rescue some of the discarded data, but the design goal here is to automate the corrections with a high throughput procedure.

The final PDBBind-Opt dataset contains 27,609 protein-ligand complexes structures and associated binding affinity data that have been updated by the workflow. The reason behind the increase in the amount of data compared to PDBBind is that we have included multiple protein-ligand complexes from the same RCSB PDB entry due to observed non-negligible structural fluctuations between chains that share sequence identity. More detailed justification is included below. In addition, considering the original PDBBind general set was further filtered to create a "refined" and "core" set based on structure quality, binding data quality, and redundancy reduction,[27] we also comply with this division of the PDBBind-Opt data to yield totals of 4,946 and 278 entries in the refined and core sets, respectively.

Compilation of BioLiP2-Opt dataset. In order to demonstrate the utility of our PDBBind-Opt workflow, we have created BioLiP2-Opt, a new and independent dataset of protein-ligand complex structures and their associated binding affinity values. The BioLiP2 database[36] is used as a starting point, which provides a sizable collection of biologically relevant protein-ligand binding entries deposited in RCSB PDB enriched with multiple annotations, including binding affinity data from various sources (PDBBind,[27] Binding MOAD,[51] BindingDB,[52,53] and manual annotation). We first downloaded the txt-formatted database from the BioLiP official website and select entries based on the following rules:

- Only entries with at least one binding affinity (K_i, K_d or IC_{50}) annotations are accepted. Entries with EC_{50} are not accepted for reasons described above. Six PDB entries (2BXG, 2I30, 3T74, 3T8G, 4H57, 6TMN) are also discarded because their binding affinities are invalid (with $K_i > 10^3$ M).

- Only entries with one valid ligand CCD associated are accepted. This excludes cases that the ligand is a polymer or multiple ligands bind to the same pocket.

- The ligand must have at least six heavy atoms. This excludes inorganic molecules like O_2, CO_2, CO_3^{2-} which is uncommon in protein-ligand binding studies.

- Entries that exist in PDBBind v2020 are not accepted. This allows users to treat BioLiP2-Opt as an external benchmark for various docking methods or SFs that have been trained on PDBBind-Opt, or it can be combined with PDBBind-Opt to create an expanded dataset.

After the qualified BioLiP2 entries were selected, the PDBBind-Opt workflow described above was used to yield 3,343 unique PDB entries along with 5,687 protein-ligand complex structures.

To characterize and validate the BioLiP2-Opt dataset, Figure 3 provides the distributions of binding affinities and drug-like properties compared to the PDBBind-Opt dataset. It is seen that the new BioLiP2-Opt dataset shares a very close set of distributions of these properties with PDBBind-Opt, especially for the binding affinities in which both datasets cover a large window with approximate 10 log units. This demonstrate the reliability of the new BioLiP2-Opt dataset as a useful resource for future SFs development, benchmarks and other structure-based drug design studies.
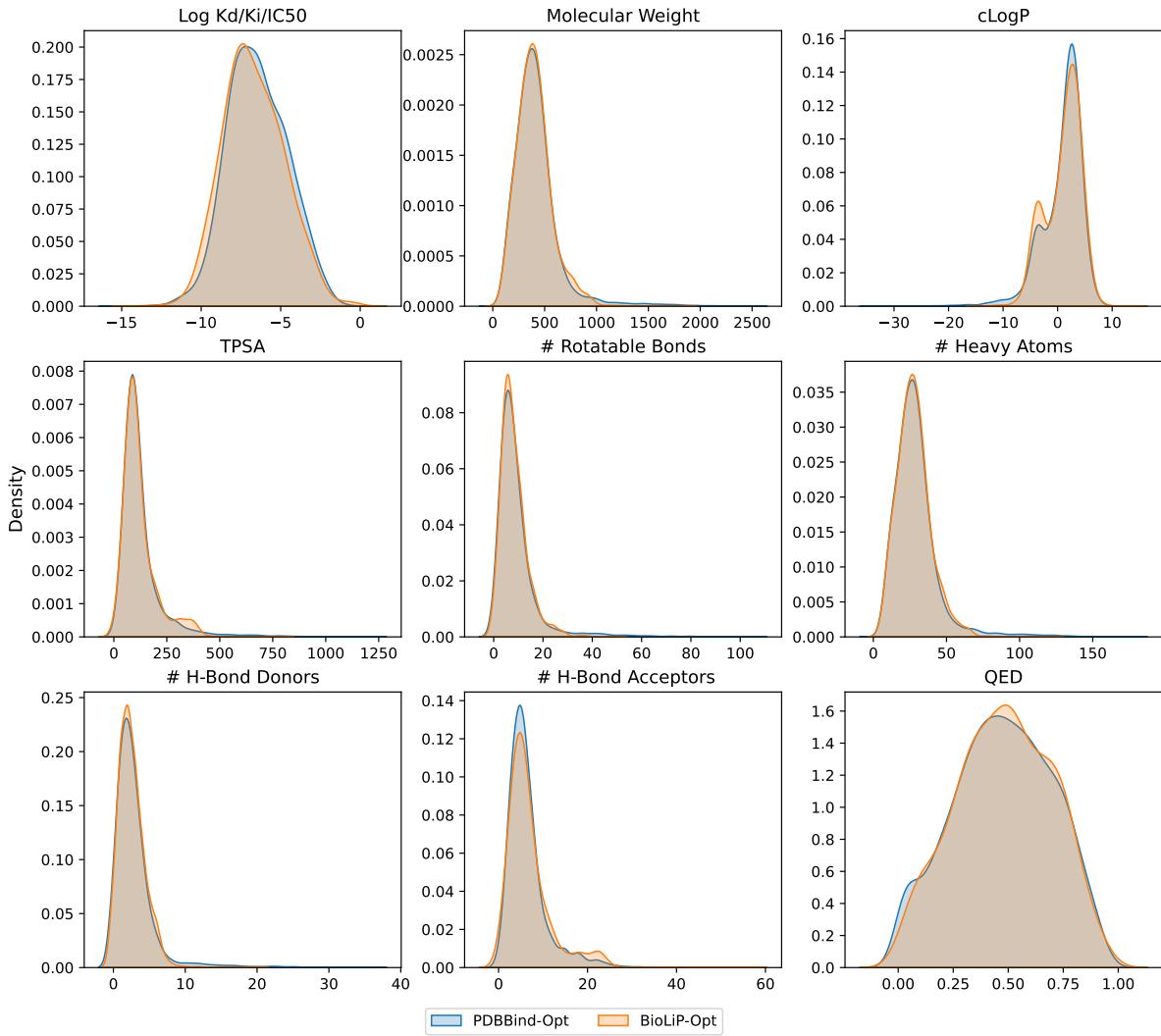
Figure 3: *Comparison of the distribution of drug-related properties of ligands and their binding affinities to proteins for PDBBind-Opt and BioLiP2-Opt.* (a) Binding Kd, Ki, and IC50 values in log units. (b) molecular weight, (c) computed log P value, (d) topological polar surface area (TPSA), (e) the number of rotatable bonds, (f) the number of heavy atoms, (g) the number of hydrogen bond donor atoms, (h)the number of hydrogen bond acceptor atoms, and (i) quantitative estimation of drug-likeness (QED) values.

Examples of refined ligand structures using PDBBind-Opt Here we provide additional examples of the fixed ligand structures obtained from the PDBBind-Opt workflow and compared with the deposited ligands in the original PDBBind dataset as provided in Figure 4. In some cases we find that some of the ligands in PDBBind are different from what was actually reported in the literature from which they were derived. For 2AXI,[54] the ligand

13

of interest should be the cyclic peptide-like inhibitor, not the sulfonic acid buffer. In other cases, the PDBBind ligand structures are incomplete or the bonding is incorrect (Figure 4a). For example, the ligand in 1ALW[55] is missing an iodine atom and in 1DY4[56] the isopropyl is falsely reported as a cyclopropyl (Figure 4b,c). This type of problem may arise from historical reasons, i.e. some structures in PDBBind were derived from older version of RCSB PDB that contained these incorrect structures. We also find that PDBBind-Opt yields ligands with better protonation/tautomer states. Two examples are 1DG9[57] and 3DJF,[58] for which PDBBind shows that the former case contains a neutral sulfonic acid and a divalent piperazine cation motif while the latter case falsely makes a guanosine-like compound positively charged (Figure 4d,e). In this case the corresponding fixed structures in PDBBind-Opt are more chemically feasible and also in line with the protonation states predicted by ChemAxon Marvin.[44]
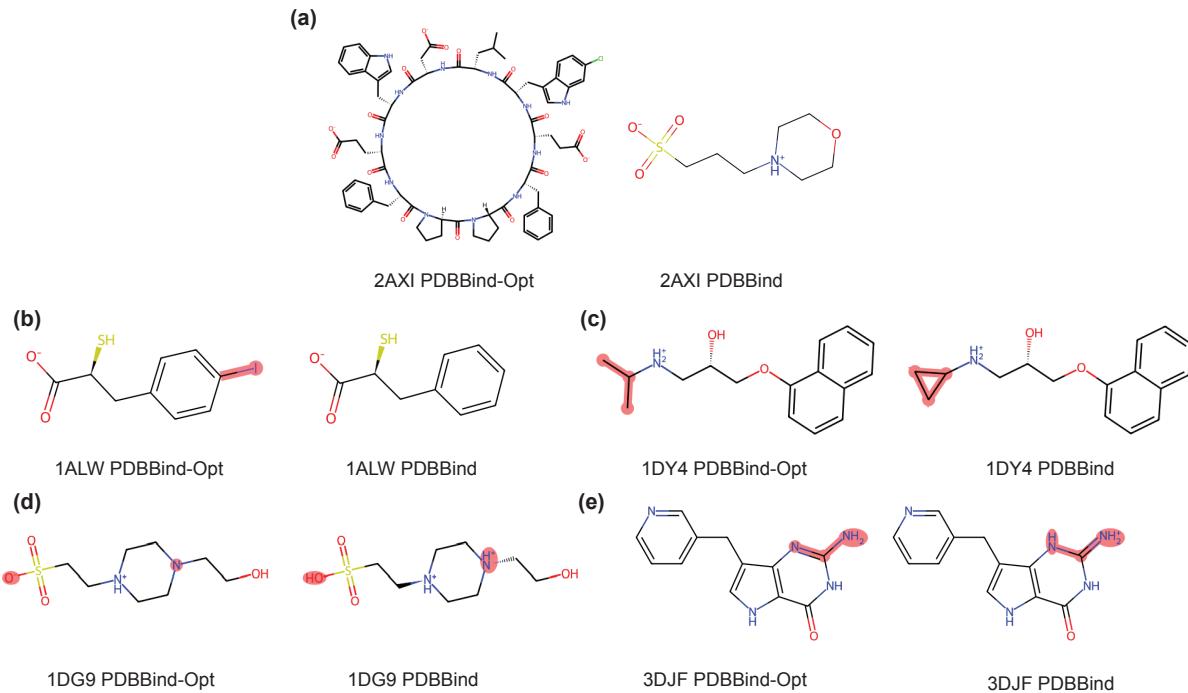

(a)

2AXI PDBBind-Opt   2AXI PDBBind

(b)

1ALW PDBBind-Opt   1ALW PDBBind

(c)

1DY4 PDBBind-Opt   1DY4 PDBBind

(d)

1DG9 PDBBind-Opt   1DG9 PDBBind

(e)

3DJF PDBBind-Opt   3DJF PDBBind

Figure 4: *Examples of corrected ligands derived from PDB-Bind-Opt compared to the original PDBBind.* (a) Wrong ligand entity reported. (b) Missing atoms. (c) Wrong bond connectivity. (d-e) Undesired protonation and tautomer states. The mol2 format ligand files in PDBBind database were used for analysis.

PDBBind provides two file formats for the ligand structure: mol2 and sdf. However, among all 19,443 entries, 45 mol2 files and 3175 sdf files cannot be processed by RDKit, a widely-used open-source cheminformatics tool. The is due to miscellaneous reasons, for instance, undesired aromaticity specification (oxygens tagged as aromatic to represent equivalent atoms in RSO_3^-, $RCOO^-$, RPO_3^{2-}) or formal charge specification (nitrogen with 4 explicit valence tagged to be neutral). PDBBind-Opt workflow naturally addresses this problem because it uses RDKit[46] to process ligand structures.
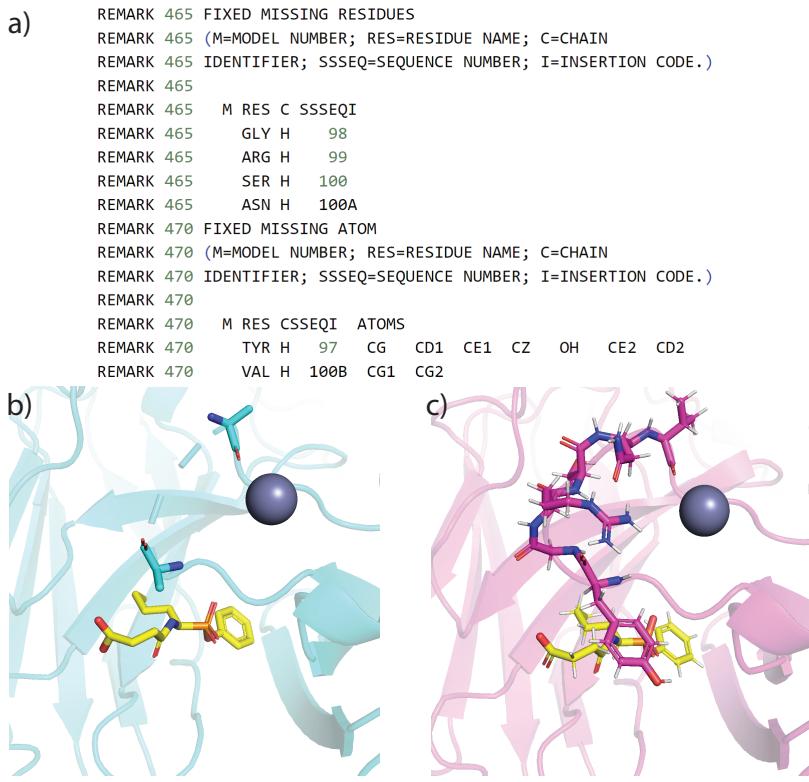

Figure 5: *Example of a corrected protein derived from PDB-Bind-Opt compared to the original PDBBind.* (a) The protein fixing metadata related to residue Y97 through V100B in the refined protein file. (b) Visualization of the original PDB entry with missing residues and atoms centralized at the region Y97-V100B close to the binding site. (c) Visualization of the refined protein structure after the protein fixing module.

Example of refined protein structures using PDBBind-Opt. With the protein fixing module, users interested in training 3D-based SFs and capturing local protein-ligand interactions would benefit from a more complete protein and binding site representation. To

demonstrate our protein fixing module, Figure 5 shows an example of protein 1A0Q[59] that has both missing atoms and missing residues around the binding site. Here, the protein fixing module first identified those missing data and fixed them based on the sequence information provided in the mmcif file header and the predefined residue templates. The reason behind using information from the mmcif header rather than the pdb "SEQRES" field is that in some of the deposited structures, missing residues are also omitted in the "SEQRES" field. As a result, an unphysical peptide bond will be placed between the start and the end of a short sequence of the missing residues, which will cause problems in training SFs. The metadata from this fixing call is stored in the refined pdb file in case users want to label the original crystal residues and repaired residues differently.

Structural variations within the same RCSB PDB entry. When compiling the PDBBind-Opt and BioLiP2-Opt datasets, we noticed that there were a moderate amount of PDB entries that contain multiple records of the same ligand of interest in the deposited structures. The reason behind a majority of such observation lies in the fact that proteins can form various quaternary structures using copies of the same chain, and ligands as binders can interact with the macromolecule at the tertiary, chain, or quaternary level. As a result, when a ligand binds to a specific chain of the protein, it will show up, for instance, two times in a PDB entry that is a homodimer of that chain. PDBBind[27] usually keeps only one randomly-selected sample of the interacting protein-ligand complex. However, since different chains in PDB are resolved separately using their electron density maps, there are still some level of non-negligible structural variations among different copies, making them valuable data sources for training SFs.

As shown in Figure 6, although the RMSD distribution between identical chains of the same entry do not show a great difference, there is a significant amount of rotamer state changes observed across these chains. Here, following the common practice in the field,[60] we used the angle cutoff of 60° to any of the four side chain torsion angles to define a switch in the rotamer states. We also provide an example using the PDB entry 3GEP[61] to show

16

the large structural differences between chain A and B. In this case, 29 out of 57 residues near the binding site are calculated to have a change in their rotameric states, including 12 residues in the free loop area (L101, S103-I113). In particular, the distance between the side chain of D107 and the ligand in chain A (blue) is smaller than 4Å, compared to chain B where the free loop is further apart from the ligand. Therefore, through this example, it is clear that including multiple records of protein-ligand interaction within the same PDB entry is necessary.
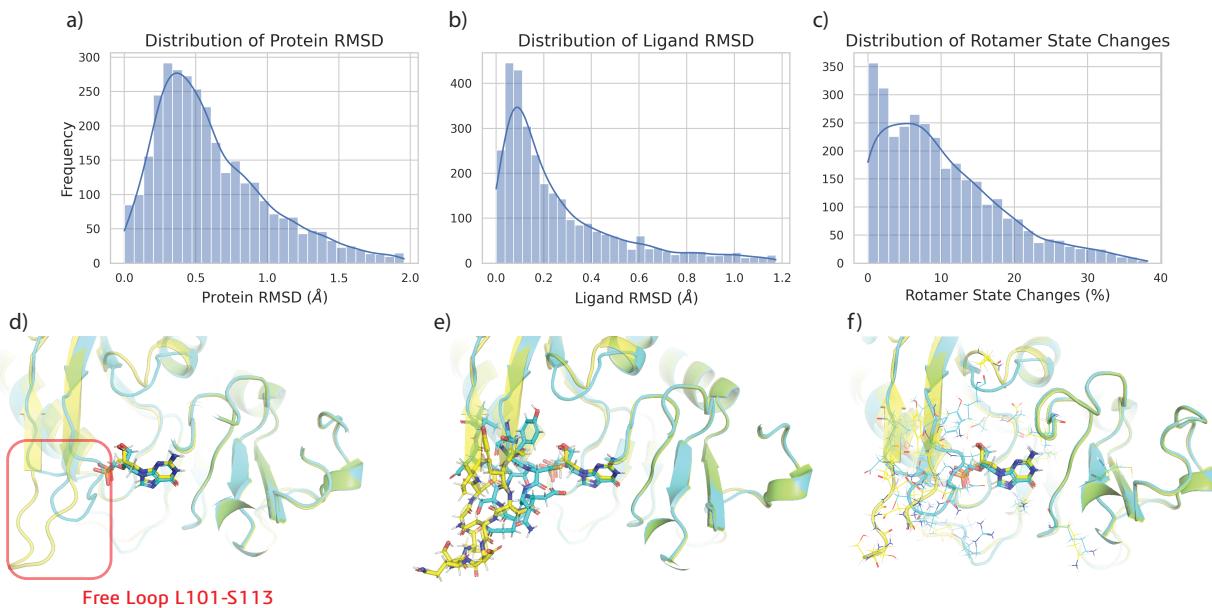

Figure 6: *Structural variations within the same RCSB PDB entry.* Top row: distribution of different structural properties for chains from the same PDB entry that have identical sequences. (a) Distribution of protein RMSD. (b) Distribution of ligand RMSD. (c) Percentage of rotamer state changes for residues around the ligand binding sites. Bottom row: visualization of changes in rotamer states between chain A (blue) and chain B (yellow) of PDB ID: 3GEP. (d) Structure overlay between two chains and their respective ligands. (e) Structural differences around the free loop regions between two chains visualized as sticks. (f) Rotamer comparisons of all 29 residues that change their states across chains.

DATA RECORDS

The PDBBind-Opt and BioLiP2-Opt dataset can be found in a Figshare repository.[62] The refined ligand structures are in sdf format and the refined protein structure are in pdb format. The metadata, including the ligand CCD code, ligand chain id, ligand residue number, PDB deposit year, resolution, binding affinity annotaions, is provided in a csv file. A table providing column names and a description of the contents of each column can also be found in the Figshare repository.

USAGE NOTES

For easy storage, the structural data in PDBBind-Opt and BioLiP-Opt is compressed to a gzipped tarball file (.tar.gz). After unzipping it, there will be a folder for each unique PDB ID, which contains one or more subfolder(s) that follows the naming convention: `PDBID_LIG_CHAIN_RESNUM`, where LIG is the ligand's name, CHAIN is the ligand chain ID and RESNUM is the ligand residue number. If the ligand is a polymer, LIG and RESNUM will be specified by its first and last residue with a hyphen, for example "ACE-DIP" and "100-103". Each subfolder contains 6 files: "*_ligand.pdb", "*_protein.pdb" are respectively ligand and protein structures directly extracted from the original PDB without any refinement; "*_hetatm.pdb" is the structure of the additives (ions, solvents, co-factors) and "*_protein_hetatm.pdb" is the protein structure with additives (no refinement); "*_ligand_refined.sdf" and "*_protein_refined.pdb" are the prepared structures with PDBBind-Opt workflow.

DATA AND CODE AVAILABILITY

All the codes for PDBBind-Opt workflow and BioLiP2-Opt dataset creation are provided in a public accessible GitHub repository: https://github.com/THGLab/PDBBind-Opt

AUTHOR CONTRIBUTIONS

Y.W., K.S., J.L. and T.H.-G. conceived the scientific direction for PDBBind-Opt workflow and BioLiP2-Opt dataset and wrote the manuscript. Y.W. and K.S. wrote the codes and prepared the dataset. All authors provided comments on the results and manuscript.

Acknowledgement

References

(1) Huang, S.-Y.; Grinter, S. Z.; Zou, X. Scoring functions and their evaluation methods for protein–ligand docking: recent advances and future directions. *Phys. Chem. Chem. Phys.* **2010**, *12*, 12899–12908.

(2) Trott, O.; Olson, A. J. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J. Comp. Chem.* **2010**, *31*, 455–461.

(3) Morris, G. M.; Huey, R.; Lindstrom, W.; Sanner, M. F.; Belew, R. K.; Goodsell, D. S.; Olson, A. J. AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. *J. Comp. Chem.* **2009**, *30*, 2785–2791.

(4) Friesner, R. A.; Banks, J. L.; Murphy, R. B.; Halgren, T. A.; Klicic, J. J.; Mainz, D. T.; Repasky, M. P.; Knoll, E. H.; Shelley, M.; Perry, J. K.; others Glide: a new approach

for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. *J. Med. Chem.* **2004**, *47*, 1739–1749.

(5) Friesner, R. A.; Murphy, R. B.; Repasky, M. P.; Frye, L. L.; Greenwood, J. R.; Halgren, T. A.; Sanschagrin, P. C.; Mainz, D. T. Extra precision glide: Docking and scoring incorporating a model of hydrophobic enclosure for protein- ligand complexes. *J. Med. Chem.* **2006**, *49*, 6177–6196.

(6) Qiu, Y. et al. Development and Benchmarking of Open Force Field v1.0.0-the Parsley Small-Molecule Force Field. *J Chem Theory Comput* **2021**, *17*, 6262–6280.

(7) Boothroyd, S. et al. Development and Benchmarking of Open Force Field 2.0.0: The Sage Small Molecule Force Field. *J. Chem. Theo. Comp.* **2023**, *19*, 3251–3275.

(8) Wang, R.; Lai, L.; Wang, S. Further development and validation of empirical scoring functions for structure-based binding affinity prediction. *Journal of computer-aided molecular design* **2002**, *16*, 11–26.

(9) Jones, G.; Willett, P.; Glen, R. C.; Leach, A. R.; Taylor, R. Development and validation of a genetic algorithm for flexible docking. *J. Mol. Bio.* **1997**, *267*, 727–748.

(10) Muegge, I. PMF scoring revisited. *J. Med. Chem.* **2006**, *49*, 5895–5902.

(11) Huang, N.; Kalyanaraman, C.; Bernacki, K.; Jacobson, M. P. Molecular mechanics methods for predicting protein–ligand binding. *Phys. Chem. Chem. Phys.* **2006**, *8*, 5166–5177.

(12) Dittrich, J.; Schmidt, D.; Pfleger, C.; Gohlke, H. Converging a knowledge-based scoring function: DrugScore2018. *J. Chem. Info. Model.* **2018**, *59*, 509–521.

(13) Ballester, P. J.; Mitchell, J. B. A machine learning approach to predicting protein–ligand binding affinity with applications to molecular docking. *Bioinformatics* **2010**, *26*, 1169–1175.

(14) Jiang, D.; Hsieh, C.-Y.; Wu, Z.; Kang, Y.; Wang, J.; Wang, E.; Liao, B.; Shen, C.; Xu, L.; Wu, J.; others Interactiongraphnet: A novel and efficient deep graph representation learning framework for accurate protein–ligand interaction predictions. *J. Med. Chem.* **2021**, *64*, 18209–18232.

(15) Moon, S.; Zhung, W.; Yang, S.; Lim, J.; Kim, W. Y. PIGNet: a physics-informed deep learning model toward generalized drug–target interaction predictions. *Chemical Science* **2022**, *13*, 3661–3673.

(16) Shen, C.; Zhang, X.; Deng, Y.; Gao, J.; Wang, D.; Xu, L.; Pan, P.; Hou, T.; Kang, Y. Boosting protein–ligand binding pose prediction and virtual screening based on residue–atom distance likelihood potential and graph transformer. *J. Med. Chem.* **2022**, *65*, 10691–10706.

(17) Ozturk, H.; Ozgur, A.; Ozkirimli, E. DeepDTA: deep drug–target binding affinity prediction. *Bioinformatics* **2018**, *34*, i821–i829.

(18) Somnath, V. R.; Bunne, C.; Krause, A. Multi-scale representation learning on proteins. *Advances in Neural Information Processing Systems* **2021**, *34*, 25244–25255.

(19) Lu, W.; Wu, Q.; Zhang, J.; Rao, J.; Li, C.; Zheng, S. Tankbind: Trigonometry-aware neural networks for drug-protein binding structure prediction. *bioRxiv* **2022**, 2022–06.

(20) Yang, Z.; Zhong, W.; Lv, Q.; Dong, T.; Yu-Chian Chen, C. Geometric interaction graph neural network for predicting protein–ligand binding affinities from 3d structures (gign). *The journal of physical chemistry letters* **2023**, *14*, 2020–2033.

(21) Wang, R.; Fang, X.; Lu, Y.; Wang, S. The PDBbind Database: Collection of Binding Affinities for Protein-Ligand Complexes with Known Three-Dimensional Structures. *Journal of Medicinal Chemistry* **2004**, *47*, 2977–2980.

(22) Wang, R.; Fang, X.; Lu, Y.; Yang, C.-Y.; Wang, S. The PDBbind Database: Methodologies and Updates. *Journal of Medicinal Chemistry* **2005**, *48*, 4111–4119.

(23) Cheng, T.; Li, X.; Li, Y.; Liu, Z.; Wang, R. Comparative Assessment of Scoring Functions on a Diverse Test Set. *Journal of Chemical Information and Modeling* **2009**, *49*, 1079–1093.

(24) Li, Y.; Liu, Z.; Li, J.; Han, L.; Liu, J.; Zhao, Z.; Wang, R. Comparative Assessment of Scoring Functions on an Updated Benchmark: 1. Compilation of the Test Set. *Journal of Chemical Information and Modeling* **2014**, *54*, 1700–1716.

(25) Li, Y.; Han, L.; Liu, Z.; Wang, R. Comparative Assessment of Scoring Functions on an Updated Benchmark: 2. Evaluation Methods and General Results. *Journal of Chemical Information and Modeling* **2014**, *54*, 1717–1736.

(26) Liu, Z.; Li, Y.; Han, L.; Li, J.; Liu, J.; Zhao, Z.; Nie, W.; Liu, Y.; Wang, R. PDB-wide collection of binding data: current status of the PDBbind database. *Bioinformatics* **2015**, *31*, 405–412.

(27) Liu, Z.; Su, M.; Han, L.; Liu, J.; Yang, Q.; Li, Y.; Wang, R. Forging the basis for developing protein–ligand interaction scoring functions. *Accounts of chemical research* **2017**, *50*, 302–309.

(28) Cheng, T.; Li, X.; Li, Y.; Liu, Z.; Wang, R. Comparative assessment of scoring functions on a diverse test set. *J. Chem. Info. Model.* **2009**, *49*, 1079–1093.

(29) Su, M.; Yang, Q.; Du, Y.; Feng, G.; Liu, Z.; Li, Y.; Wang, R. Comparative assessment of scoring functions: the CASF-2016 update. *J. Chem. Info. Model.* **2018**, *59*, 895–913.

(30) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Research* **2000**, *28*, 235–242.

(31) Zhu, K.; Borrelli, K. W.; Greenwood, J. R.; Day, T.; Abel, R.; Farid, R. S.; Harder, E. Docking covalent inhibitors: a parameter free approach to pose prediction and scoring. *Journal of chemical information and modeling* **2014**, *54*, 1932–1940.

(32) Bianco, G.; Forli, S.; Goodsell, D. S.; Olson, A. J. Covalent docking using autodock: Two-point attractor and flexible side chain methods. *Protein Science* **2016**, *25*, 295–301.

(33) Delbart, F.; Brams, M.; Gruss, F.; Noppen, S.; Peigneur, S.; Boland, S.; Chaltin, P.; Brandao-Neto, J.; Von Delft, F.; Touw, W. G.; Joosten, R. P.; Liekens, S.; Tytgat, J.; Ulens, C. An allosteric binding site of the $\alpha 7$ nicotinic acetylcholine receptor revealed in a humanized acetylcholine-binding protein. *Journal of Biological Chemistry* **2018**, *293*, 2534–2545.

(34) Rosner, K. E. et al. The discovery of novel tartrate-based TNF-α converting enzyme (TACE) inhibitors. *Bioorganic & Medicinal Chemistry Letters* **2010**, *20*, 1189–1193.

(35) Yang, J.; Roy, A.; Zhang, Y. BioLiP: a semi-manually curated database for biologically relevant ligand-protein interactions. *Nucleic Acids Research* **2013**, *41*, D1096–1103.

(36) Zhang, C.; Zhang, X.; Freddolino, P. L.; Zhang, Y. BioLiP2: an updated structure database for biologically relevant ligand–protein interactions. *Nucleic Acids Research* **2023**, *52*, D404–D412.

(37) Li, J.; Guan, X.; Zhang, O.; Sun, K.; Wang, Y.; Bagni, D.; Head-Gordon, T. Leak Proof PDBBind: A Reorganized Dataset of Protein-Ligand Complexes for More Generalizable Binding Affinity Prediction. *ArXiv* **2024**, arXiv:2308.09639v2.

(38) PDBFixer. `https://github.com/openmm/pdbfixer`, accessed on Oct 29, 2024.

(39) Liu, Z. H.; Tsanai, M.; Zhang, O.; Forman-Kay, J.; Head-Gordon, T. Computational

Methods to Investigate Intrinsically Disordered Proteins and their Complexes. 2024; https://arxiv.org/abs/2409.02240.

(40) Gall, T. L.; Romero, P. R.; Cortese, M. S.; Uversky, V. N.; Dunker, A. K. Intrinsic disorder in the protein data bank. *Journal of Biomolecular structure and dynamics* **2007**, *24*, 325–341.

(41) O'Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R. Open Babel: An open chemical toolbox. *Journal of cheminformatics* **2011**, *3*, 1–14.

(42) Luo, W.; Zhou, G.; Zhu, Z.; Yuan, Y.; Ke, G.; Wei, Z.; Gao, Z.; Zheng, H. Bridging Machine Learning and Thermodynamics for Accurate p K a Prediction. *JACS Au* **2024**, *4*, 3451–3465.

(43) Johnston, R. C.; Yao, K.; Kaplan, Z.; Chelliah, M.; Leswing, K.; Seekins, S.; Watts, S.; Calkins, D.; Chief Elk, J.; Jerome, S. V.; others Epik: p K a and Protonation State Prediction through Machine Learning. *Journal of chemical theory and computation* **2023**, *19*, 2380–2388.

(44) Prediction of dissociation constant using microconstants. https://docs.chemaxon.com/display/docs/attachments/attachments_1814016_1_Prediction_of_dissociation_constant_using_microconstants.pdf, accessed on Oct 31, 2024.

(45) Ropp, P. J.; Kaminsky, J. C.; Yablonski, S.; Durrant, J. D. Dimorphite-DL: an open-source program for enumerating the ionization states of drug-like small molecules. *Journal of Cheminformatics* **2019**, *11*, 1–8.

(46) Landrum, G.; others RDKit: A software suite for cheminformatics, computational chemistry, and predictive modeling. *Greg Landrum* **2013**, *8*.

(47) Maier, J. A.; Martinez, C.; Kasavajhala, K.; Wickstrom, L.; Hauser, K. E.; Simmerling, C. ff14SB: improving the accuracy of protein side chain and backbone parameters from ff99SB. *Journal of chemical theory and computation* **2015**, *11*, 3696–3713.

(48) Eastman, P. et al. OpenMM 8: Molecular Dynamics Simulation with Machine Learning Potentials. *The Journal of Physical Chemistry B* **2024**, *128*, 109–116, PMID: 38154096.

(49) Fermi, E. *Thermodynamics*; Courier Corporation, 2012.

(50) MarÉchal, E. Measuring bioactivity: KI, IC50 and EC50. *Chemogenomics and Chemical Genetics: A User's Introduction for Biologists, Chemists Informaticians* **2011**, 55–65.

(51) Benson, M. L.; Smith, R. D.; Khazanov, N. A.; Dimcheff, B.; Beaver, J.; Dresslar, P.; Nerothin, J.; Carlson, H. A. Binding MOAD, a high-quality protein-ligand database. *Nucleic Acids Research* **2008**, *36*, D674–678.

(52) Liu, T.; Lin, Y.; Wen, X.; Jorissen, R. N.; Gilson, M. K. BindingDB: a web-accessible database of experimentally determined protein–ligand binding affinities. *Nucleic acids research* **2007**, *35*, D198–D201.

(53) Gilson, M. K.; Liu, T.; Baitaluk, M.; Nicola, G.; Hwang, L.; Chong, J. BindingDB in 2015: A public database for medicinal chemistry, computational chemistry and systems pharmacology. *Nucleic Acids Research* **2016**, *44*, D1045–D1053.

(54) Fasan, R.; Dias, R. L. A.; Moehle, K.; Zerbe, O.; Obrecht, D.; Mittl, P. R. E.; Grütter, M. G.; Robinson, J. A. Structure–Activity Studies in a Family of β-Hairpin Protein Epitope Mimetic Inhibitors of the p53–HDM2 Protein–Protein Interaction. *ChemBioChem* **2006**, *7*, 515–526.

(55) Lin, G.-d.; Chattopadhyay, D.; Maki, M.; Wang, K. K. W.; Carson, M.; Jin, L.; Yuen, P.-w.; Takano, E.; Hatanaka, M.; DeLucas, L. J.; Narayana, S. V. Crystal structure of calcium bound domain VI of calpain at 1.9 Å resolution and its role in enzyme

assembly, regulation, and inhibitor binding. *Nature Structural Biology* **1997**, *4*, 539–547.

(56) Ståhlberg, J.; Henriksson, H.; Divne, C.; Isaksson, R.; Pettersson, G.; Johansson, G.; Jones, T. Structural basis for enantiomer binding and separation of a common β-blocker: crystal structure of cellobiohydrolase Cel7A with bound (S)-propranolol at 1.9 Å resolution. *Journal of Molecular Biology* **2001**, *305*, 79–93.

(57) Zhang, M.; Zhou, M.; Van Etten, R. L.; Stauffacher, C. V. Crystal Structure of Bovine Low Molecular Weight Phosphotyrosyl Phosphatase Complexed with the Transition State Analog Vanadate ˙. *Biochemistry* **1997**, *36*, 15–23.

(58) Castilho, M. S.; Postigo, M. P.; Pereira, H. M.; Oliva, G.; Andricopulo, A. D. Structural basis for selective inhibition of purine nucleoside phosphorylase from Schistosoma mansoni: Kinetic and structural studies. *Bioorganic & Medicinal Chemistry* **2010**, *18*, 1421–1427.

(59) Buchbinder, J. L.; Stephenson, R. C.; Scanlan, T. S.; Fletterick, R. J. A comparison of the crystallographic structures of two catalytic antibodies with esterase activity11Edited by I. A. Wilson. *Journal of Molecular Biology* **1998**, *282*, 1033–1041.

(60) Shapovalov, M. V.; Dunbrack, R. L. A smoothed backbone-dependent rotamer library for proteins derived from adaptive kernel density estimates and regressions. *Structure* **2011**, *19*, 844–858.

(61) Keough, D. T.; Hocková, D.; Holý, A.; Naesens, L. M. J.; Skinner-Adams, T. S.; Jersey, J. d.; Guddat, L. W. Inhibition of Hypoxanthine-Guanine Phosphoribosyltransferase by Acyclic Nucleoside Phosphonates: A New Class of Antimalarial Therapeutics. *Journal of Medicinal Chemistry* **2009**, *52*, 4391–4399, PMID: 19527031.

(62) Wang, Y.; Sun, K.; Li, J.; Guan, X.; Zhang, O.; Bagni, D.; Head-Gordon, T. PDBBind Optimization to Create a High-Quality Protein-Ligand Binding Dataset for

Binding Affinity Prediction. 2024; `https://figshare.com/collections/PDBBind_Optimization_to_Create_a_High-Quality_Protein-Ligand_Binding_Dataset_for_Binding_Affinity_Prediction/7520133/1`.