Digital Discovery

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal | View Issue

Cite this: *Digital Discovery*, 2025, **4**, 1209

A workflow to create a high-quality protein–ligand binding dataset for training, validation, and prediction tasks†

Yingze Wang, ‡[a] Kunyang Sun,‡[a] Jie Li, [a] Xingyi Guan, [a] Oufan Zhang,[a] Dorian Bagni,[a] Yang Zhang,[def] Heather A. Carlson[g] and Teresa Head-Gordon *[abc]

Received 5th November 2024
Accepted 25th March 2025

DOI: 10.1039/d4dd00357h

rsc.li/digitaldiscovery

Development of scoring functions (SFs) used to predict protein–ligand binding energies requires high-quality 3D structures and binding assay data for training and testing their parameters. In this work, we show that one of the widely-used datasets, PDBbind, suffers from several common structural artifacts of both proteins and ligands, which may compromise the accuracy, reliability, and generalizability of the resulting SFs. Therefore, we have developed a series of algorithms organized in a semi-automated workflow, HiQBind-WF, that curates non-covalent protein–ligand datasets to fix these problems. We also used this workflow to create an independent data set, HiQBind, by matching binding free energies from various sources including BioLiP, Binding MOAD and Binding DB with co-crystalized ligand–protein complexes from the PDB. The resulting HiQBind workflow and 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.

1 Introduction

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. Various benchmarks based on PDBbind, such as CASF (Comparative Assessment of Scoring Functions) series,[28–31] CSAR (Community Structure Activity Resource) 2010 (ref. 32) and PDBbind-blind-2013 (ref. 33) have been proposed to assess the scoring power, ranking power, docking power and screening power of various SFs.

PDBbind has been an invaluable resource to the biomolecular community during its two-decade development, but a significant portion of the PDBbind dataset contains structural errors, statistical anomalies, and a sub-optimal organization of protein–ligand classes that can limit SF training and validation.[34,35] 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. Another concern in regards PDBbind is that the data processing procedure is neither open-sourced nor automated, potentially relying on individual groups needing to introduce their own manual intervention that may lead to inconsistencies. Furthermore, the PDBbind data curation process became more problematic in 2021 when PDBbind ceased to be freely available for data curated after 2020, which limits access and hinders the development and validation of new scoring functions (and other additional uses).

[a]*Kenneth S. Pitzer Theory Center and Department of Chemistry, University of California, Berkeley, CA, 94720, USA*
[b]*Department of Bioengineering, University of California, Berkeley, CA, 94720, USA*
[c]*Departments of Bioengineering and Chemical and Biomolecular Engineering, University of California, Berkeley, CA, 94720, USA*
[d]*Department of Computer Science, School of Computing, National University of Singapore, 117417, Singapore*
[e]*Cancer Science Institute of Singapore, National University of Singapore, 117599, Singapore*
[f]*Department of Biochemistry, Yong Loo Lin School of Medicine, National University of Singapore, 117596, Singapore*
[g]*Odyssey Therapeutics Inc., 1350 Highland Dr, Ann Arbor, MI, 48108, USA*

† Electronic supplementary information (ESI) available: Details of the data formats and usage notes are provided. See DOI: https://doi.org/10.1039/d4dd00357h

‡ Authors contributed equally.

Fortunately, other curation efforts have created alternative protein–ligand structural and/or binding datasets that have increased the size and comprehensiveness of available data for drug discovery efforts. BindingDB is a database containing 2.9 million binding measurements spanning 1.3 million compounds for thousands of protein targets, which are curated from the literature and patents.[36–38] Binding MOAD is a curated database of 41 409 protein–ligand structural complexes, with binding affinity data available for 15 223 (37%) of them; Binding MOAD's curation involved extracting high-quality structures from the PDB and finding associated binding data from publications with the aid of an NLP-based annotation tool.[39–42] BioLiP is a large database of over 900 000 biologically-relevant protein–ligand interactions curated from the PDB, and enriched with various functional annotations, including Enzyme Commission numbers, Gene Ontology terms, catalytic sites, and binding affinities from Binding MOAD, BindingDB, as well as manual surveys.[43,44] Other related datasets that focuses more on the geometries of proteins and ligands, including PLINDER[45] and DockGen,[46] contain an expanded set of protein–ligand structural complexes but do not have annotations of binding affinity data. However, in general, these curation efforts have largely focused on increasing the size and comprehensiveness of protein–ligand data, rather than increasing the quality and reliability of the data themselves. Therefore, there is a pressing need for an open-source and systematic workflow to prepare protein–ligand binding datasets with well-defined binding affinity annotations and higher-quality structures in order to foster greater reproducibility, transparency, and accessibility.

In this work, we introduce HiQBind-WF, a workflow of algorithms for data cleaning and structural preparation that creates a curated dataset of high-quality, non-covalent protein–ligand complex structures with binding affinity annotations. This workflow contains several modules: (1) a curating procedure that rejects ligands covalently bonded to proteins, ligands with rarely-occurring 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 and, when necessary, add missing atoms to all chains involved in the protein–ligand binding; (4) a structure refinement module to simultaneously add hydrogens to both proteins and ligands in their complex state, as opposed to the current practice in PDBbind that completes the hydrogen chemistry for protein and ligand independently. The motivation for adding this hydrogen growth module is that although many SFs only take heavy atoms into consideration, future physics-based SFs could potentially benefit from explicit hydrogens to better model intermolecular interactions such as hydrogen bonding.

We utilized this workflow to optimize PDBbind v2020 and compared the processed structures. Analysis of the structural differences between the same PDB entry demonstrated that HiQBind-WF is able to correct for various observed structural imperfections. Further, to illustrate the applicability of the HiQBind-WF, we created HiQBind, a new dataset with high-quality protein–ligand binding structures and affinities by processing PDB entries included in BioLiP2 and Binding MOAD associated with binding affinities drawn from BindingDB. The HiQBind dataset includes >18 000 unique PDB entries and >30 000 protein–ligand complex structures. We also confirmed that HiQBind shares similar properties with existing datasets like PDBbind, demonstrating its feasibility to be used for developing and validating SFs and other structure-based drug-design tools. The HiQBind-WF and HiQBind dataset are provided open-source to foster transparency and sustainability as new data appears, in order to maintain this important resource for the biology and drug discovery communities.

2 Methods

The flowchart of HiQBind-WF is illustrated in Fig. 1. We start by downloading the pdb and mmcif formats directly from the RCSB PDB[47] for supplied entries. 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, and curate these categories as follows.

We define three classes of ligand(s) for any given protein–ligand complex structure: (1) any residue will be identified as a ligand if its name matches the Chemical Component Dictionary (CCD) code deposited in given reference datasets (PDBbind, BioLiP or Binding MOAD). Ligands identified in this manner are referred to as "small molecules". (2) Otherwise, chains in the original PDB file that are less than 20 residues but more than one residue as ligands will be selected as ligands. For example, PDBbind entries that contain patterns such as "*-mer," or symbols like "–", "&" or "+", MOAD entries with more than one CCD in the name column of its csv-formatted dataset, and BioLiP entries with Ligand CCD column to be "peptide", "dna" or "rna". These ligands are typically polypeptides, oligosaccharides, or oligonucleotides, collectively referred to as "polymers". (3) For each identified ligand, we label any biopolymer chains within 10 Å as the associated protein structure. Then, for each protein structure, we labeled residues specified by the "HETATM" record in the pdb file within 4 Å 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 splitting for the ligand categories and their associated protein–ligand complexes, we define an additional set of downselect filters, including some borrowed from the processing protocols of LP-PDBbind.[34] The purpose of these filters are to exclude protein–ligand complex structures that specifically can interfere with training of SFs, with eliminations that would meet any of the following criteria:

• 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. This is because covalent binding inherently is different from non-covalent binding which does not involve bond breaking, and thus requires special treatment in any SFs. Those covalent binder entries are included in the ESI† which

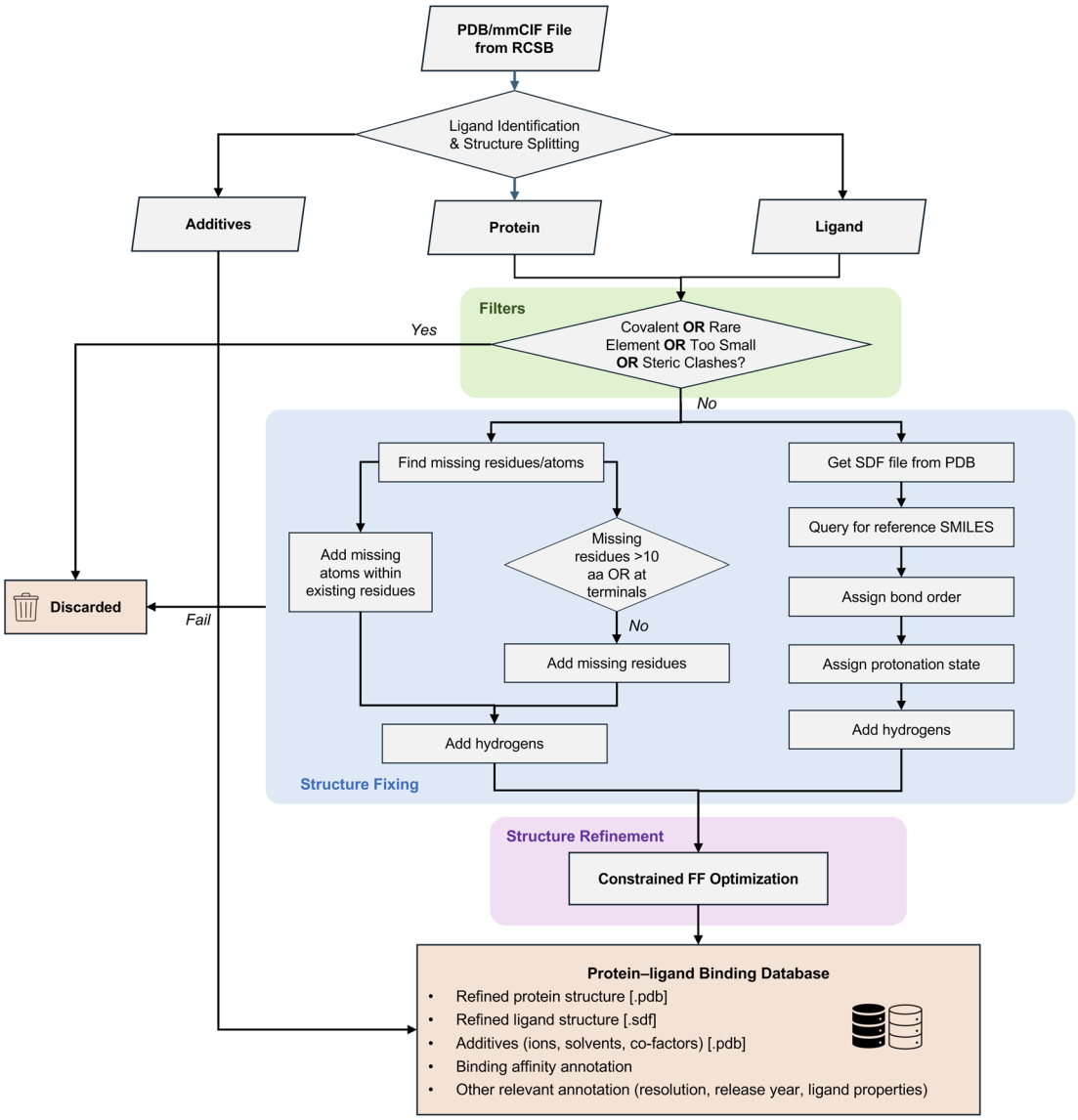

Fig. 1 Schematic representation of the semi-automated HiQBind-WF to refine protein–ligand binding structures. HiQBind-WF downloads the pdb and mmCIF files from the RCSB PDB,[47] followed by splitting each structure into three components—ligand, protein, and additives. A series of filters are then applied to remove covalent binders, ligands with rare elements, very small ligands, and complexes exhibiting steric clashes. Subsequently, the protein structure is fixed by adding missing atoms and residues (ProteinFixer) and the ligand structure is fixed by correcting bond orders, protonation states, and aromaticity (LigandFixer). Finally, the fixed protein and ligand structures are recombined and subjected to a constrained energy minimization to resolve potential unreasonable structures and to refine hydrogen positions.

may be helpful as a separate curation of the data or may be accessible from CovBinderInPDB.[48]

• Rare element filter: excludes ligands 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 giving data sparsity for these ligands. These ligand entries are also included in the ESI† which may be helpful as a separate curation of the data.

• Small ligand filter: excludes ligands containing less than 4 heavy atoms, which includes small inorganic binders like O_2, NH_3, CO_2, NO_2, N_3^-, which are beyond the scope of common protein–ligand binding studies.

• Steric clashes filter: 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.

For protein–ligand complexes that pass these filters, two structure-fixing modules are implemented separately for proteins and ligands. In the ProteinFixer module, we first use

the sequence information from the mmcif file header to detect missing atoms and residues. Then, for missing residues or missing atoms within an existing residue, PDBFixer[49] (version 1.9) is used to add them, except when the missing residues are longer than 10 amino acids or are located at the sequence terminals. Adding missing atoms to protein structure is essential near binding sites 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 correct structural ensemble are sampled during simulations. However, long missing segments or missing terminus residues in crystal structure are often attributable to intrinsically disordered regions (IDR),[50] domains that are not expressed in the samples for crystallization, or his-tags introduced in the protein purification process.[51] If far enough removed from the ligand binding site(s), we regard it safe to skip modeling these residues explicitly. Hence, we leave these regions in their original form and in themselves do not define a criterion for being discarded in the final dataset. The final step of the protein-fixing module is to add hydrogen atoms at pH = 7.4 with PDBFixer. At this pH, the protonation state assignment of titratable side-chains obeys the following rules: 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 whether the HID or HIE variant (the hydrogen is added to Nδ or Nε, respectively) is selected will be based on which one forms a better hydrogen bond, which is the default behavior of PDBFixer.

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.[52] Since explicit atom connections may not be 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 point, 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.

After a correct structure is obtained, protonation states are assigned to the ligand. 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.[53–55] 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 the dimophite_dl package[56] which can be found in the Github repository.

One thing that 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[47]) 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 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 bonds. 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 it's beyond the scope of the workflow. 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 state assignment, and added hydrogens in the ligand-fixing module are all performed with RDKit[57] (version 2024.03.4).

The last part of the HiQBind-WF structure preparation 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[58] is used for the protein and OpenFF-2.1.0 (ref. 7) together with Gasteiger charges[59] are used for the ligand. 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. Additionally, the hydrogen-bonding network between the ligand and protein is also optimized in this process. The constrained energy minimization was performed with OpenMM 8.1.1 (ref. 60) by setting masses of all constrained atoms to zero.

The binding affinity in terms of ΔG is directly related to the dissociation coefficient K_d or K_i through the standard relationship $\Delta G = RT \ln(K_{d/i})$.[61] However, a large portion of the data in the binding datasets 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.[62] Furthermore, the IC_{50} values for the same protein–ligand complex can vary up to over order of magnitude in different assays,[63] and some deposited binding data 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, IC_{50} and EC_{50}.

3 Results

Fig. 2 illustrates the common problems that arise in the training and tests sets for protein–ligand interactions and associated binding assays when developing a scoring function using various curated databases. Some of the structural imperfections are inherited from the original RCSB Protein Data Bank (PDB)[47] dataset, such as 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 such as shown in Fig. 2a, which requires special methods to account for the covalent bond formulation,[64,65] and should remain distinct from protein–ligand complexes that are formed from non-covalent interactions only. Fig. 2b illustrates an example from the PDBbind refined set, 5OUH,[66] which is a non-covalent binder that exhibits a severe atomic clash with the protein. Similarly, for 3KMC[67] the chlorobenzene portion of the ligand is absent from the crystal structure as seen in Fig. 2c. This again highlights the need for the community to have a free open-source tool to curate high-quality protein–ligand structures in a reproducible way.

3.1 The HiQBind workflow

We applied HiQBind-WF to refine the structures in the publicly available PDBbind v2020 dataset.[27] While we are not able to publish this optimized PDBbind dataset because the user's agreement of PDBbind prohibits any distribution of any derivative dataset, we can report some general statistics of the workflow and provide examples of structural fixes of the protein and ligand data. However, users can reproduce an optimized PDBbind data set using HiQBind-WF following step-by-step instructions in our Github repository.

Of the original 19 443 unique PDB entries for proteins with ligands in PDBbind v2020, 1330 entries were discarded by the filters and 2452 entries were discarded because they were unable to pass the structure fixing and refinement modules. Our final optimized PDBbind dataset contains 15 661 unique PDB IDs and 27 757 protein–ligand complexes structures. In addition, considering that 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] the optimized PDBbind data yields totals of 4969 and 279 entries in the refined and core sets, respectively. Finally, the associated 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, IC_{50} and EC_{50}.

3.1.1 Example of refined ligand structures using HiQBind-WF. Here, we also provide examples of the fixed ligand structures obtained from HiQBind-WF and compared with the deposited ligands in the original PDBbind dataset as provided in Fig. 3. 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,[68] the ligand 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 (Fig. 3a). For example, the ligand in 1ALW[69] is missing an iodine atom and in 1DY4 (ref. 70) the isopropyl is falsely reported as a cyclopropyl (Fig. 3b and c). This type of problem may arise for historical reasons, *i.e.* some structures in PDBbind were derived from

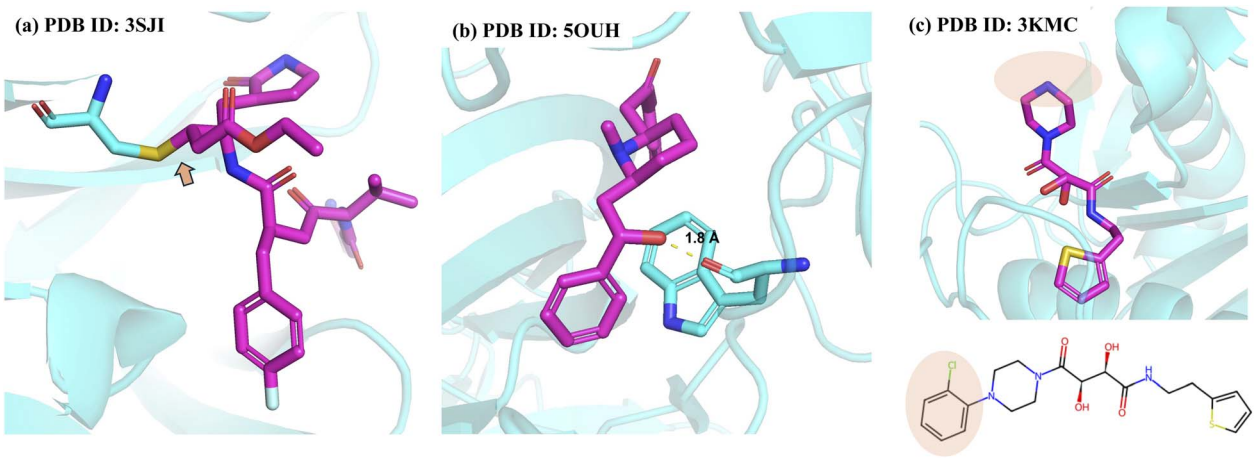

Fig. 2 Common structural imperfections in PDBbind dataset. (a) Covalent binders. The ligand is covalently bound to cysteine with a Michael addition reaction. (b) Steric clashes with the distance between the clashing atoms being only 1.8 Å. (c) Missing atoms.

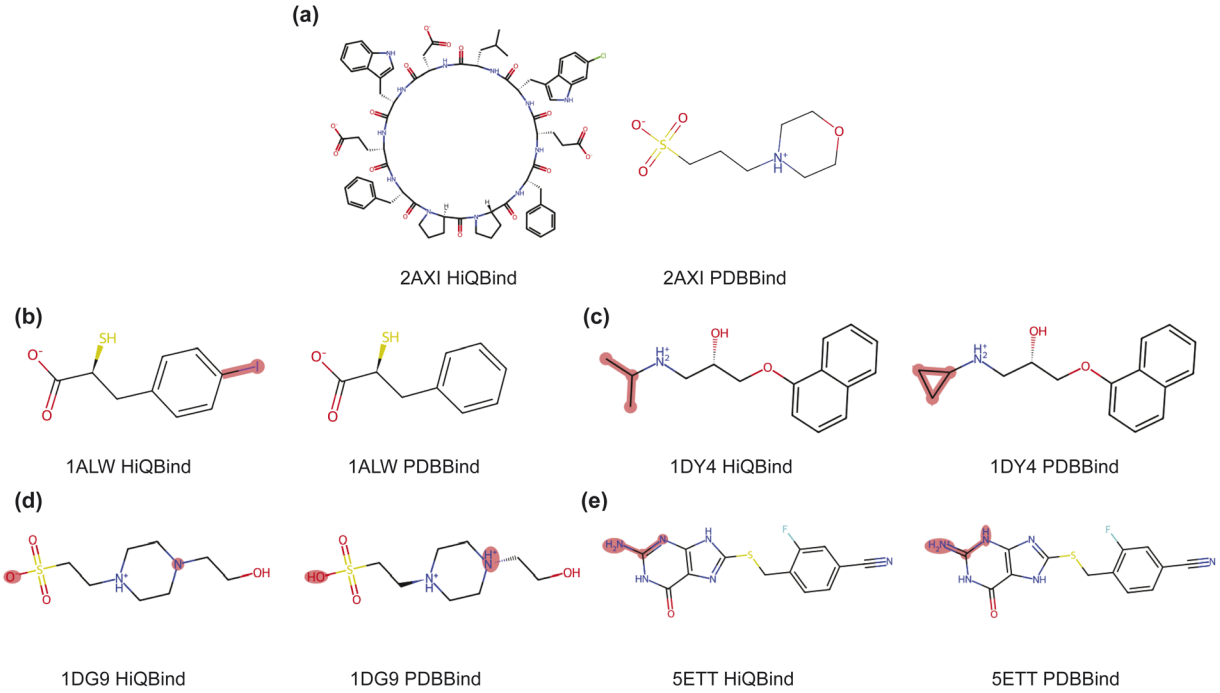

Fig. 3 Examples of corrected ligands by HiQBind-WF compared to the original PDBbind. (a) Wrong ligand entity reported. (b) Missing atoms. (c) Wrong bond connectivity. (d) and (e) Undesired protonation and tautomer states. The mol2 format ligand files in PDBbind database were used for analysis.

older version of RCSB PDB that contained these incorrect structures. We also find that HiQBind-WF yields ligands with better protonation/tautomer states. Two examples are 1DG9 (ref. 71) and 5ETT,[72] 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 (Fig. 3d and e). In this case, the fixed ligand structures are more chemically feasible and also in line with the protonation states predicted by ChemAxon Marvin.[55]

HiQBind-WF also fixes a small but practical problem in PDBbind. 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[57] (version 2024.03.4), a widely-used open-source cheminformatics tool. This may be due to the fact that these files are prepared by some other software and their sdf specifications are not compatible with RDKit. Examples are 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). HiQBind-WF naturally addresses this technical problem because it uses RDKit[57] to process ligand structures.

3.1.2 Example of refined protein structures using HiQBind-WF. 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, Fig. 4 shows an example of protein 1A0Q[73] 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 repair residues differently.

3.2 Creation of the HiQBind dataset

In order to further demonstrate the utility of HiQBind-WF, we have created a new dataset of high-quality, non-covalent protein–ligand complex structures and their associated binding affinity values. To prepare the HiQBind dataset, we used two biologically relevant protein–ligand datasets as a starting point: BioLiP2 (ref. 44) and Binding MOAD.[41] We downloaded the txt-formatted BioLiP database from its official website and csv-formatted dataset MOAD from its Github repository and entries with at least one reported binding affinity (K_i, K_d, IC_{50} or EC_{50}) data were selected. BioLiP2 itself provides a sizable collection of protein–ligand entries deposited in RCSB PDB and enriched with multiple annotations, including binding affinity data from various sources including Binding MOAD,[41] BindingDB,[36,37] and manual annotation. Although BioLiP2 encompasses much of Binding MOAD, we still found additional entries from Binding MOAD that we also include in

a)

```
REMARK 465 FIXED MISSING RESIDUES
REMARK 465 (M=MODEL NUMBER; RES=RESIDUE NAME; C=CHAIN
REMARK 465 IDENTIFIER; SSSEQ=SEQUENCE NUMBER; I=INSERTION CODE.)
REMARK 465
REMARK 465   M RES C SSSEQI
REMARK 465     GLY H    98
REMARK 465     ARG H    99
REMARK 465     SER H   100
REMARK 465     ASN H   100A
REMARK 470 FIXED MISSING ATOM
REMARK 470 (M=MODEL NUMBER; RES=RESIDUE NAME; C=CHAIN
REMARK 470 IDENTIFIER; SSSEQ=SEQUENCE NUMBER; I=INSERTION CODE.)
REMARK 470
REMARK 470   M RES CSSEQI  ATOMS
REMARK 470     TYR H   97   CG   CD1  CE1  CZ   OH   CE2  CD2
REMARK 470     VAL H  100B  CG1  CG2
```

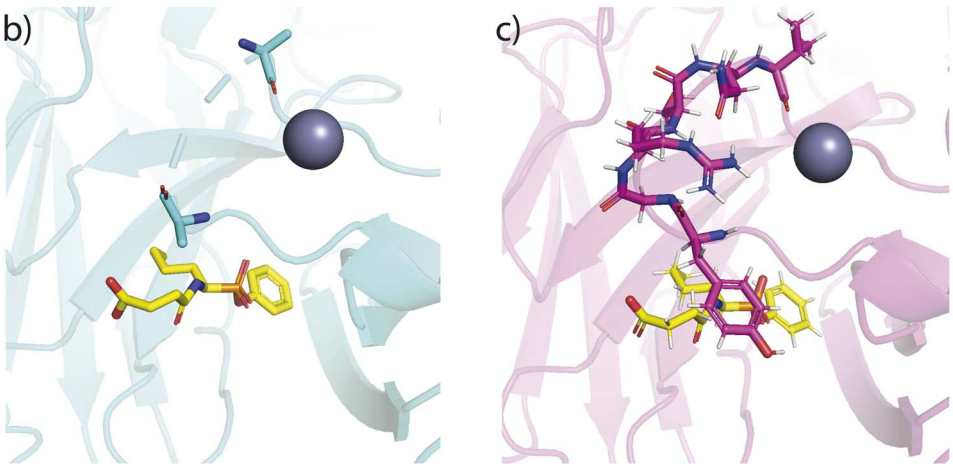

Fig. 4 Example of a refined protein structure derived from HiQBind-WF compared to the original PDBbind. (a) The protein fixing metadata related to residue Y97 through V100B in the refined protein file. (b) Visualization of the original PDB entry with missing residues and atoms centralized at the region Y97-V100B close to the binding site. (c) Visualization of the refined protein structure after the protein-fixing module.

our new dataset. Over this entire merged set seven PDB entries (2BXG, 2I30, 3T74, 3T8G, 4H57, 6TMN and 7JWN) were discarded because their binding affinities are invalid (with $K_i > 10^3$ M), and all entries as part of the publicly available PDBbind v2020 dataset are not included. In total, 20 349 unique PDB entries with reported binding affinities were obtained.

We then applied HiQBind-WF to process all these starting entries, yielding 18 160 unique PDB entries. For the 2189 entries that failed to pass HiQBind-WF, 761 of them are discarded by the filters and the remaining 1428 entries are due to the failure of structure fixing and refinement modules (Table S4†). A large portion of the discarded datapoints are "polymers" for which it is hard to verify their structural correctness because of the difficulty in obtaining a reliable reference SMILES string which is more suitable to small molecules. 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 as much as possible.

At the same time we retain 32 275 protein–ligand complex structures. The reasons 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 because: (1) multiple conformers or stereoisomers can contribute to the binding (Fig. 5a); (2) the same ligand can bind to different protein pockets (Fig. 5b); (3) there is more than one ligand with a reported binding affinity (Fig. 5c); 131 PDB entries are of this reason; and (4) for structures containing homo-multimers, structural fluctuations between chains that share sequence identity are non-negligible.

A moderate amount of PDB entries contain multiple records of the same ligand of interest in the deposited structures. The reason 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 or quaternary level. For example, when a ligand binds to a specific chain of a homodimer protein, two PDB entries are present. 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.

Although the overall protein and ligand RMSD distributions between identical chains of the same entry do not show a great difference (Fig. 5d and e), there is a significant amount of side-chain rotamer state changes observed for different chains as

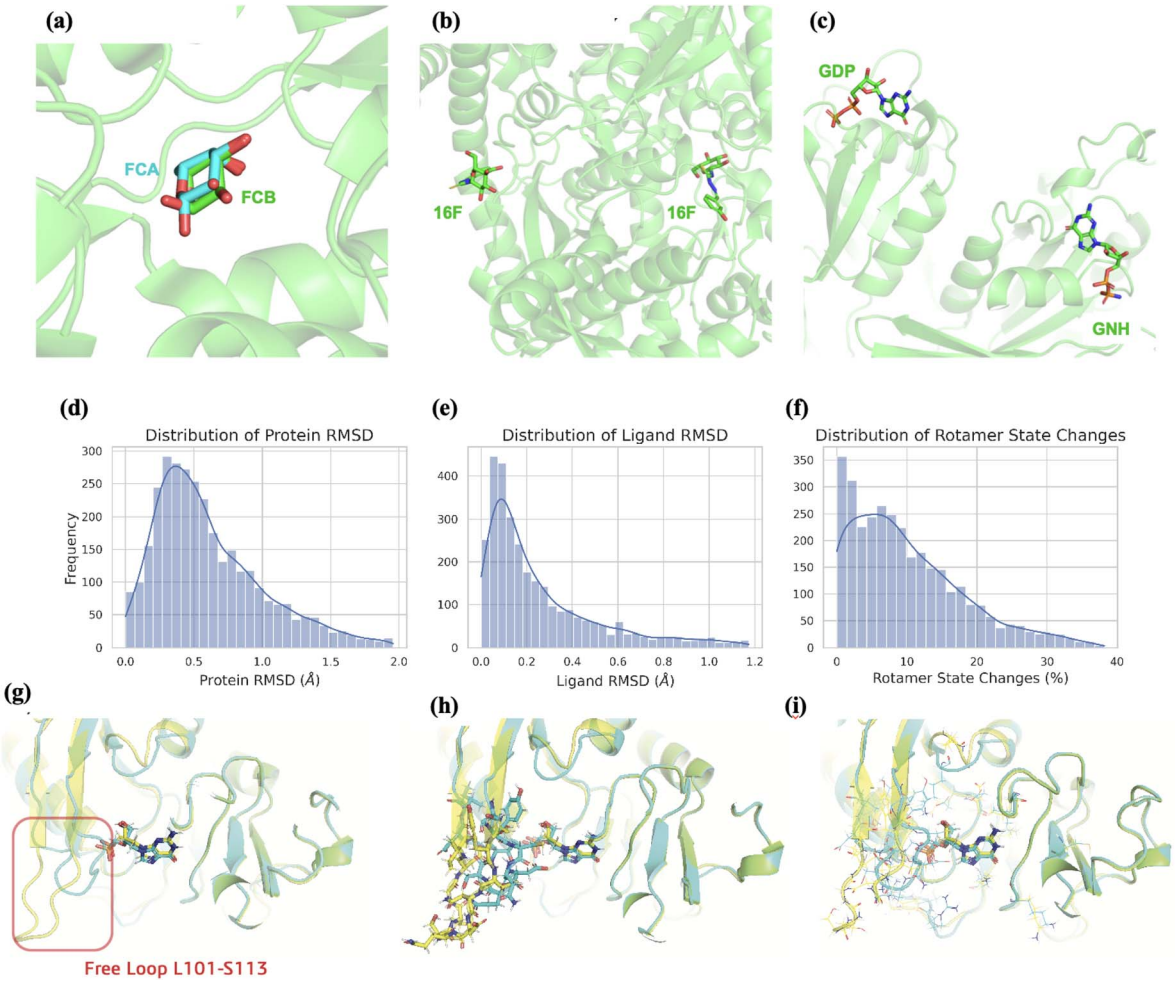

Fig. 5 Structural variations and more than one protein–ligand complexes in the same PDB entry. (a) Two stereoisomers (FCA and FCB) contribute to the binding for 1ABF. (b) The same ligand (16F) binds to two different pockets in 3MRV (c) Two different ligands (GDP) and (GNH) bind to the protein in different pockets for 1A4R. (d) Distribution of protein RMSD. (e) Distribution of ligand RMSD. (f) 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. (g) Structure overlay between two chains and their respective ligands. (h) Structural differences around the free loop regions between two chains visualized as sticks. (i) Rotamer comparisons of all 29 residues that change their states across chains.

shown in Fig. 5f. Here, following the common practice in the field,[74] we used the angle cutoff of 60° to any of the four side-chain torsion angles to define a switch in the rotamer states. To illustrate, the protein chains A and B for PDB entry 3GEP[75] in Fig. 5g and h shows that 29 out of 57 residues near the binding site have a change in their side-chain 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 4 Å, compared to chain B where the free loop is further from the ligand. Therefore, including multiple records of protein–ligand interactions with the same PDB entry can be informative and beneficial.

To characterize and validate the HiQBind dataset, Fig. 6 provides the distributions of binding affinities and drug-like properties compared to the PDBbind dataset. It is seen that the new HiQBind dataset shares a very close set of distributions of drug-like properties with PDBbind, especially for the binding affinities in which both datasets cover a large window of approximately 10 log units. We also noticed that HiQBind dataset is a bit more druglike (QED score) with smaller ligands having fewer rotatable bonds and better $c\log P$/hydrogen-bonding properties. This demonstrates the feasibility of the new HiQBind dataset as a useful resource for future SFs development, benchmarks and other structure-based drug design studies. It is also important to note that, since no standardized method exists for further filtering and data splitting, it is up to the users decide how to perform these additional operations. For example, one might filter out NMR structures and entries with IC_{50} or EC_{50} values, as done in the PDBbind refined set,[24] or split the data based on time,[33,45] sequence similarity,[34] ECOD classifications,[45,46] or protein–ligand interaction profiles.[45] In the ESI,† we also provided analysis upon the time distribution of PDB entries in HiQBind as well as its overlapping with PDBbind v2020 (Fig. S1†). This shall benefit users who want to do time-based splitting or

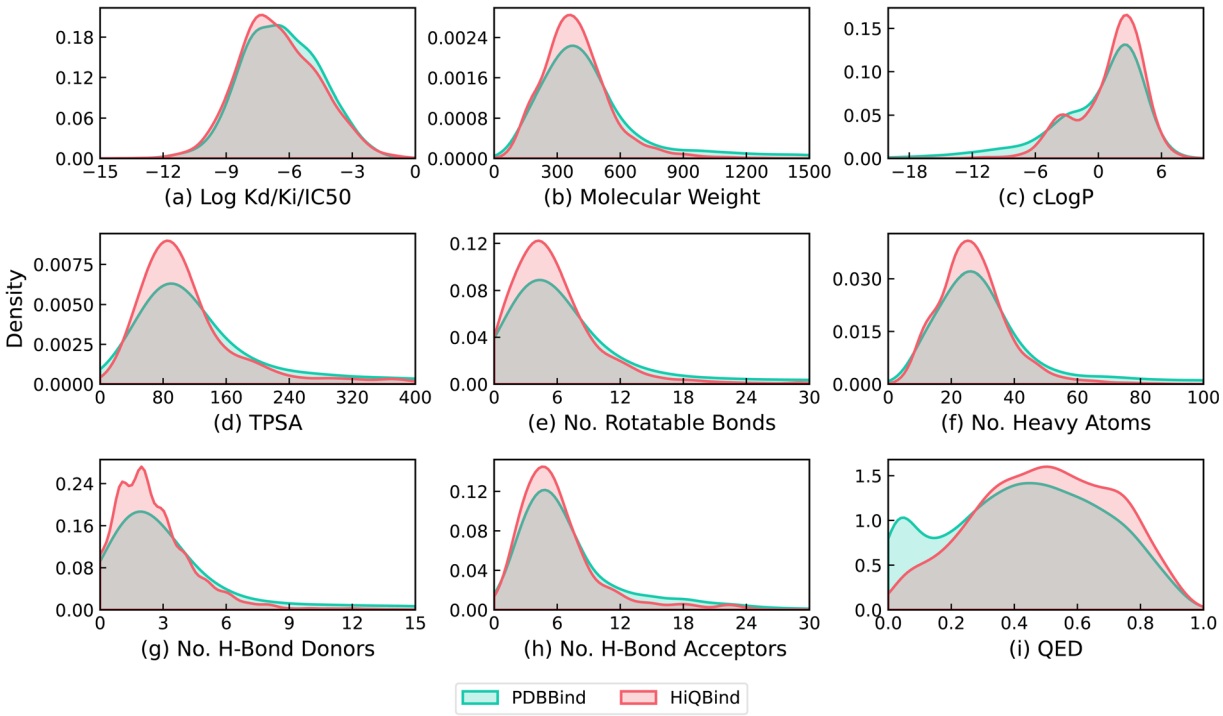

Fig. 6 Comparison of the distribution of drug-related properties of ligands and their binding affinities between PDBbind and HiQBind. (a) Binding K_d, K_i, and IC_{50} values in log units. (b) Molecular weight, (c) computed log *P* value, (d) topological polar surface area (TPSA), (e) the number of rotatable 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.

create independent test set for those models that have already been trained on PDBbind v2020.

4 Conclusions

Many physics-based and machine-learned scoring functions used to predict protein–ligand binding affinities rely on powerful databases such as PDBbind,[21–27] BioLiP2,[43,44] Binding MOAD,[41] and related databases such as Binding DB,[36,37] Plinder,[45] and Dockgen.[46] While central to the biomolecular and drug discovery communities, all data curation efforts require ongoing quality-control efforts. In fact, the latest PDBbind version hosted on the PDBbind+ website have performed some corrections, but PDBbind data curated after v2021 has been commercialized such that it is only accessible to paid users, and there is no published literature describing their workflow. Hence, we are unable to make a fair comparison between the quality of their generated data compared to that presented here. Therefore, we believe there is a need for the community to have a free open-source tool to curate high-quality protein–ligand structures in a reproducible way.

We have developed an optimization workflow, HiQBind-WF that aims to improve the structural integrity in a semi-automated way and produce high-quality structures with binding affinity annotations. We compared PDBbind v2020 to the structures processed with HiQBind-WF. Differences between the complexes highlighted the strength of our workflow in assigning correct bond orders, protonation states, and protein structure refinements. We also used this workflow to prepare HiQBind, a newly compiled dataset based on BioLiP2 (ref. 44) and Binding MOAD[41] that offers high-quality, non-covalent protein–ligand complexes with binding-affinity data. HiQBind provides more than 30 000 protein–ligand structures spanning over 18 000 unique PDB entries and is feasible to be deployed in the development of scoring functions or force fields or related activities.

As an open source effort, we believe that HiQBind-WF provides a sustainable framework for continuously updating and refining protein–ligand binding datasets for drug discovery, by meeting scientific goals of ensuring transparency and reproducibility. We also envision that structure-based drug design studies can benefit from the new HiQBind data that has no overlap with PDBbind, and thus reporting evaluation metrics upon this new dataset that will become a common practice as part of future computational modeling efforts.

Data and code availability

All the codes for HiQBind-WF workflow and HiQBind dataset creation are provided in a public accessible GitHub repository: **https://github.com/THGLab/HiQBind** under MIT License. The associated DOI is: **https://doi.org/10.5281/zenodo.14903380**. The HiQBind dataset, including the protein–ligand structures, metadata information (binding affinity annotations, release year, resolution, ligand name, protein name, protein UniProt ID and various ligand properties) is publicly available in fig share: **https://doi.org/10.6084/m9.figshare.27430305**.

Author contributions

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

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

We thank Michael Gilson for helpful discussions. This work was supported by National Institute of Allergy and Infectious Disease grant U19-AI171954. This research used computational resources of the National Energy Research Scientific Computing, a DOE Office of Science User Facility supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.

References

1 S. Y. Huang, S. Z. Grinter and X. Zou, Scoring functions and their evaluation methods for protein–ligand docking: recent advances and future directions, *Phys. Chem. Chem. Phys.*, 2010, **12**(40), 12899–12908.
2 O. Trott and A. J. Olson, AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading, *J. Comput. Chem.*, 2010, **31**(2), 455–461.
3 G. M. Morris, R. Huey, W. Lindstrom, M. F. Sanner, R. K. Belew, D. S. Goodsell, *et al.*, AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility, *J. Comput. Chem.*, 2009, **30**(16), 2785–2791.
4 R. A. Friesner, J. L. Banks, R. B. Murphy, T. A. Halgren, J. J. Klicic, D. T. Mainz, *et al.*, Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy, *J. Med. Chem.*, 2004, **47**(7), 1739–1749.
5 R. A. Friesner, R. B. Murphy, M. P. Repasky, L. L. Frye, J. R. Greenwood, T. A. Halgren, *et al.*, Extra precision glide: Docking and scoring incorporating a model of hydrophobic enclosure for protein–ligand complexes, *J. Med. Chem.*, 2006, **49**(21), 6177–6196.
6 Y. Qiu, D. G. A. Smith, S. Boothroyd, H. Jang, D. F. Hahn, J. Wagner, *et al.*, Development and Benchmarking of Open Force Field v1.0.0-the Parsley Small-Molecule Force Field, *J. Chem. Theory Comput.*, 2021, **17**(10), 6262–6280.
7 S. Boothroyd, P. K. Behara, O. C. Madin, D. F. Hahn, H. Jang, V. Gapsys, *et al.*, Development and Benchmarking of Open Force Field 2.0.0: The Sage Small Molecule Force Field, *J. Chem. Theory Comput.*, 2023, **19**(11), 3251–3275, DOI: **10.1021/acs.jctc.3c00039**.
8 R. Wang, L. Lai and S. Wang, Further development and validation of empirical scoring functions for structure-based binding affinity prediction, *J. Comput.-Aided Mol. Des.*, 2002, **16**, 11–26.
9 G. Jones, P. Willett, R. C. Glen, A. R. Leach and R. Taylor, Development and validation of a genetic algorithm for flexible docking, *J. Mol. Biol.*, 1997, **267**(3), 727–748.
10 I. Muegge, PMF scoring revisited, *J. Med. Chem.*, 2006, **49**(20), 5895–5902.
11 N. Huang, C. Kalyanaraman, K. Bernacki and M. P. Jacobson, Molecular mechanics methods for predicting protein–ligand binding, *Phys. Chem. Chem. Phys.*, 2006, **8**(44), 5166–5177.
12 J. Dittrich, D. Schmidt, C. Pfleger and H. Gohlke, Converging a knowledge-based scoring function: DrugScore2018, *J. Chem. Inf. Model.*, 2018, **59**(1), 509–521.
13 P. J. Ballester and J. B. Mitchell, A machine learning approach to predicting protein–ligand binding affinity with applications to molecular docking, *Bioinformatics*, 2010, **26**(9), 1169–1175.
14 D. Jiang, C. Y. Hsieh, Z. Wu, Y. Kang, J. Wang, E. Wang, *et al.*, Interaction graphnet: a novel and efficient deep graph representation learning framework for accurate protein–ligand interaction predictions, *J. Med. Chem.*, 2021, **64**(24), 18209–18232.
15 S. Moon, W. Zhung, S. Yang, J. Lim and W. Y. Kim, PIGNet: a physics-informed deep learning model toward generalized drug–target interaction predictions, *Chem. Sci.*, 2022, **13**(13), 3661–3673.
16 C. Shen, X. Zhang, Y. Deng, J. Gao, D. Wang, L. Xu, *et al.*, Boosting protein–ligand binding pose prediction and virtual screening based on residue–atom distance likelihood potential and graph transformer, *J. Med. Chem.*, 2022, **65**(15), 10691–10706.
17 H. Ozturk, A. Ozgur and E. Ozkirimli, DeepDTA: deep drug–target binding affinity prediction, *Bioinformatics*, 2018, **34**(17), i821–i829, DOI: **10.1093/bioinformatics/bty593**.
18 V. R. Somnath, C. Bunne and A. Krause, Multi-scale representation learning on proteins, *Adv. Neural Inf. Process. Syst.*, 2021, **34**, 25244–25255.
19 W. Lu, Q. Wu, J. Zhang, J. Rao, C. Li and S. Zheng, Tankbind: trigonometry-aware neural networks for drug-protein binding structure prediction, *bioRxiv*, 2022, 2022–2060.
20 Z. Yang, W. Zhong, Q. Lv, T. Dong and C. C. Yu-Chian, Geometric interaction graph neural network for predicting protein–ligand binding affinities from 3d structures (gign), *J. Phys. Chem. Lett.*, 2023, **14**(8), 2020–2033.
21 R. Wang, X. Fang, Y. Lu and S. Wang, The PDBbind Database: Collection of Binding Affinities for Protein–Ligand Complexes with Known Three-Dimensional Structures, *J. Med. Chem.*, 2004, **47**(12), 2977–2980. Available from: **https://pubs.acs.org/doi/10.1021/jm030580l**.
22 R. Wang, X. Fang, Y. Lu, C. Y. Yang and S. Wang, The PDBbind Database: Methodologies and Updates, *J. Med. Chem.*, 2005, **48**(12), 4111–4119. Available from: **https://pubs.acs.org/doi/10.1021/jm048957q**.
23 T. Cheng, X. Li, Y. Li, Z. Liu and R. Wang, Comparative Assessment of Scoring Functions on a Diverse Test Set, *J.*

Chem. Inf. Model., 2009, **49**(4), 1079–1093. Available from: https://pubs.acs.org/doi/10.1021/ci9000053.

24 Y. Li, Z. Liu, J. Li, L. Han, J. Liu, Z. Zhao, *et al.*, Comparative Assessment of Scoring Functions on an Updated Benchmark: 1. Compilation of the Test Set, *J. Chem. Inf. Model.*, 2014, **54**(6), 1700–1716. Available from: https://pubs.acs.org/doi/10.1021/ci500080q.

25 Y. Li, L. Han, Z. Liu and R. Wang, Comparative Assessment of Scoring Functions on an Updated Benchmark: 2. Evaluation Methods and General Results, *J. Chem. Inf. Model.*, 2014, **54**(6), 1717–1736. Available from: https://pubs.acs.org/doi/10.1021/ci500081m.

26 Z. Liu, Y. Li, L. Han, J. Li, J. Liu, Z. Zhao, *et al.*, PDB-wide collection of binding data: current status of the PDBbind database, *Bioinformatics*, 2015, **31**(3), 405–412.

27 Z. Liu, M. Su, L. Han, J. Liu, Q. Yang, Y. Li, *et al.*, Forging the basis for developing protein–ligand interaction scoring functions, *Acc. Chem. Res.*, 2017, **50**(2), 302–309.

28 M. Su, Q. Yang, Y. Du, G. Feng, Z. Liu, Y. Li, *et al.*, Comparative assessment of scoring functions: the CASF-2016 update, *J. Chem. Inf. Model.*, 2018, **59**(2), 895–913.

29 T. Cheng, X. Li, Y. Li, Z. Liu and R. Wang, Comparative assessment of scoring functions on a diverse test set, *J. Chem. Inf. Model.*, 2009, **49**(4), 1079–1093.

30 Y. Li, Z. Liu, J. Li, L. Han, J. Liu, Z. Zhao, *et al.*, Comparative assessment of scoring functions on an updated benchmark: 1. Compilation of the test set, *J. Chem. Inf. Model.*, 2014, **54**(6), 1700–1716.

31 Y. Li, L. Han, Z. Liu and R. Wang, Comparative assessment of scoring functions on an updated benchmark: 2. Evaluation methods and general results, *J. Chem. Inf. Model.*, 2014, **54**(6), 1717–1736.

32 Jr J. B. Dunbar, R. D. Smith, C. Y. Yang, P. M. U. Ung, K. W. Lexa, N. A. Khazanov, *et al.*, CSAR benchmark exercise of 2010: selection of the protein–ligand complexes, *J. Chem. Inf. Model.*, 2011, **51**(9), 2036–2046.

33 H. Li, K. S. Leung, M. H. Wong and P. J. Ballester, Improving AutoDock Vina using random forest: the growing accuracy of binding affinity prediction by the effective exploitation of larger data sets, *Mol. Inf.*, 2015, **34**(2–3), 115–126.

34 J. Li, X. Guan, O. Zhang, K. Sun, Y. Wang, D. Bagni, *et al.*, Leak Proof PDBBind: A Reorganized Dataset of Protein–Ligand Complexes for More Generalizable Binding Affinity Prediction, *arXiv*, 2024, preprint, arXiv:2308.09639v2, DOI: 10.48550/arXiv.2308.09639.

35 H. Stärk, O. Ganea, L. Pattanaik, R. Barzilay and T. Jaakkola, Equibind: Geometric deep learning for drug binding structure prediction, in *International conference on machine learning*, PMLR, 2022. pp. 20503–20521.

36 T. Liu, Y. Lin, X. Wen, R. N. Jorissen and M. K. Gilson, BindingDB: a web-accessible database of experimentally determined protein–ligand binding affinities, *Nucleic Acids Res.*, 2007, **35**(1), D198–D201.

37 M. K. Gilson, T. Liu, M. Baitaluk, G. Nicola, L. Hwang and J. Chong, Binding DB in 2015: a public database for medicinal chemistry, computational chemistry and systems pharmacology, *Nucleic Acids Res.*, 2016, **44**(D1), D1045–D1053, DOI: 10.1093/nar/gkv1072.

38 T. Liu, L. Hwang, S. K. Burley, C. I. Nitsche, C. Southan, W. P. Walters, *et al.*, BindingDB in 2024: a FAIR knowledge base of protein-small molecule binding data, *Nucleic Acids Res.*, 2025, **53**(D1), D1633–D1644.

39 L. Hu, M. L. Benson, R. D. Smith, M. G. Lerner and H. A. Carlson, Binding MOAD (mother of all databases), *Proteins: Struct., Funct., Bioinf.*, 2005, **60**(3), 333–340.

40 A. Ahmed, R. D. Smith, J. J. Clark, Jr J. B. Dunbar and H. A. Carlson, Recent improvements to Binding MOAD: a resource for protein–ligand binding affinities and structures, *Nucleic Acids Res.*, 2015, **43**(D1), D465–D469.

41 S. Wagle, R. D. Smith, I. I. I. A. J. Dominic, D. DasGupta, S. K. Tripathi and H. A. Carlson, Sunsetting binding MOAD with its last data update and the addition of 3D-ligand polypharmacology tools, *Sci. Rep.*, 2023, **13**(1), 3008.

42 M. L. Benson, R. D. Smith, N. A. Khazanov, B. Dimcheff, J. Beaver, P. Dresslar, *et al.*, Binding MOAD, a high-quality protein–ligand database, *Nucleic Acids Res.*, 2008, **36**, D674–D678.

43 J. Yang, A. Roy and Y. Zhang, BioLiP: a semi-manually curated database for biologically relevant ligand-protein interactions, *Nucleic Acids Res.*, 2013, **41**, D1096–D1103.

44 C. Zhang, X. Zhang, P. L. Freddolino and Y. Zhang, BioLiP2: an updated structure database for biologically relevant ligand–protein interactions, *Nucleic Acids Res.*, 2023, **52**, D404–D412, DOI: 10.1093/nar/gkad630.

45 J. Durairaj, Y. Adeshina, Z. Cao, X. Zhang, V. Oleinikovas, T. Duignan, *et al.*, PLINDER: The protein–ligand interactions dataset and evaluation resource, *bioRxiv*, 2024, 2024.

46 G. Corso, A. Deng, N. Polizzi, R. Barzilay and T. Jaakkola, The Discovery of Binding Modes Requires Rethinking Docking Generalization, in: *NeurIPS 2023 Generative AI and Biology (GenBio) Workshop*, 2023.

47 H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, *et al.*, The Protein Data Bank, *Nucleic Acids Res.*, 2000, **28**(1), 235–242, DOI: 10.1093/nar/28.1.235.

48 X. K. Guo and Y. Zhang, CovBinderInPDB: A Structure-Based Covalent Binder Database, *J. Chem. Inf. Model.*, 2022, **62**(23), 6057–6068.

49 *PDBFixer*, https://github.com/openmm/pdbfixer, accessed on Oct 29, 2024.

50 Z. H. Liu, M. Tsanai, O. Zhang, J. Forman-Kay and T. Head-Gordon, Computational Methods to Investigate Intrinsically Disordered Proteins and their Complexes, *arXiv*, 2024, preprint, arXiv:2409.02240, DOI: 10.48550/arXiv.2409.02240.

51 T. L. Gall, P. R. Romero, M. S. Cortese, V. N. Uversky and A. K. Dunker, Intrinsic disorder in the protein data bank, *J. Biomol. Struct. Dyn.*, 2007, **24**(4), 325–341.

52 N. M. O'Boyle, M. Banck, C. A. James, C. Morley, T. Vandermeersch and G. R. Hutchison, Open Babel: An open chemical toolbox, *J. Cheminf.*, 2011, **3**, 1–14.

53 W. Luo, G. Zhou, Z. Zhu, Y. Yuan, G. Ke, Z. Wei, *et al.*, Bridging Machine Learning and Thermodynamics for Accurate p K a Prediction, *JACS Au*, 2024, **4**(9), 3451–3465.

54 R. C. Johnston, K. Yao, Z. Kaplan, M. Chelliah, K. Leswing, S. Seekins, *et al.*, Epik: pK_a and Protonation State Prediction through Machine Learning, *J. Chem. Theory Comput.*, 2023, **19**(8), 2380–2388.

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

56 P. J. Ropp, J. C. Kaminsky, S. Yablonski and J. D. Durrant, Dimorphite-DL: an open-source program for enumerating the ionization states of drug-like small molecules, *J. Cheminf.*, 2019, **11**, 1–8.

57 G. Landrum, *et al.*, *RDKit: a software suite for cheminformatics, computational chemistry, and predictive modeling*, Greg Landrum, 2013, vol. 8.

58 J. A. Maier, C. Martinez, K. Kasavajhala, L. Wickstrom, K. E. Hauser and C. Simmerling, ff14SB: improving the accuracy of protein side chain and backbone parameters from ff99SB, *J. Chem. Theory Comput.*, 2015, **11**(8), 3696–3713.

59 J. Gasteiger and M. Marsili, A new model for calculating atomic charges in molecules, *Tetrahedron Lett.*, 1978, **19**, 3181–3184, available from: **https://api.semanticscholar.org/CorpusID:97016098**.

60 P. Eastman, R. Galvelis, R. P. Peláez, C. R. A. Abreu, S. E. Farr, E. Gallicchio, *et al.*, OpenMM 8: Molecular Dynamics Simulation with Machine Learning Potentials, *J. Phys. Chem. B*, 2024, **128**(1), 109–116, DOI: **10.1021/acs.jpcb.3c06662**.

61 E. Fermi, *Thermodynamics*, Courier Corporation, 2012.

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

63 G. A. Ross, C. Lu, G. Scarabelli, S. K. Albanese, E. Houang, R. Abel, *et al.*, The maximal and current accuracy of rigorous protein–ligand binding free energy calculations, *Commun. Chem.*, 2023, **6**(1), 222.

64 K. Zhu, K. W. Borrelli, J. R. Greenwood, T. Day, R. Abel, R. S. Farid, *et al.*, Docking covalent inhibitors: a parameter free approach to pose prediction and scoring, *J. Chem. Inf. Model.*, 2014, **54**(7), 1932–1940.

65 G. Bianco, S. Forli, D. S. Goodsell and A. J. Olson, Covalent docking using autodock: Two-point attractor and flexible side chain methods, *Protein Sci.*, 2016, **25**(1), 295–301.

66 F. Delbart, M. Brams, F. Gruss, S. Noppen, S. Peigneur, S. Boland, *et al.*, An allosteric binding site of the α7 nicotinic acetylcholine receptor revealed in a humanized acetylcholine-binding protein, *J. Biol. Chem.*, 2018, **293**(7), 2534–2545. Available from: **https://linkinghub.elsevier.com/retrieve/pii/S0021925820317142**.

67 K. E. Rosner, Z. Guo, P. Orth, G. W. Shipps, D. B. Belanger, T. Y. Chan, *et al.*, The discovery of novel tartrate-based TNF-α converting enzyme (TACE) inhibitors, *Bioorg. Med. Chem. Lett.*, 2010, **20**(3), 1189–1193. Available from: **https://linkinghub.elsevier.com/retrieve/pii/S0960894X09017053**.

68 R. Fasan, R. L. A. Dias, K. Moehle, O. Zerbe, D. Obrecht, P. R. E. Mittl, *et al.*, Structure–Activity Studies in a Family of β-Hairpin Protein Epitope Mimetic Inhibitors of the p53–HDM2 Protein–Protein Interaction, *ChemBioChem*, 2006, **7**(3), 515–526. Available from: **https://chemistry-europe.onlinelibrary.wiley.com/doi/10.1002/cbic.200500452**.

69 G. d. Lin, D. Chattopadhyay, M. Maki, K. K. W. Wang, M. Carson, L. Jin, *et al.*, Crystal structure of calcium bound domain VI of calpain at 1.9 Å resolution and its role in enzyme assembly, regulation, and inhibitor binding, *Nat. Struct. Biol.*, 1997, **4**(7), 539–547. Available from: **https://www.nature.com/doifinder/10.1038/nsb0797-539**.

70 J. Ståhlberg, H. Henriksson, C. Divne, R. Isaksson, G. Pettersson, G. Johansson, *et al.*, 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, *J. Mol. Biol.*, 2001, **305**(1), 79–93. Available from: **https://linkinghub.elsevier.com/retrieve/pii/S0022283600942372**.

71 M. Zhang, M. Zhou, R. L. Van Etten and C. V. Stauffacher, Crystal Structure of Bovine Low Molecular Weight Phosphotyrosyl Phosphatase Complexed with the Transition State Analog Vanadate, *Biochemistry*, 1997, **36**(1), 15–23. Available from: **https://pubs.acs.org/doi/10.1021/bi961804n**.

72 M. L. Dennis, N. P. Pitcher, M. D. Lee, A. J. DeBono, Z. C. Wang, J. R. Harjani, *et al.*, Structural Basis for the Selective Binding of Inhibitors to 6-Hydroxymethyl-7,8-dihydropterin pyrophosphokinase from *Staphylococcus aureus* and *Escherichia coli*, *J. Med. Chem.*, 2016, **59**(11), 5248–5263.

73 J. L. Buchbinder, R. C. Stephenson, T. S. Scanlan and R. J. Fletterick, A comparison of the crystallographic structures of two catalytic antibodies with esterase activity, *J. Mol. Biol.*, 1998, **282**(5), 1033–1041. Available from: **https://www.sciencedirect.com/science/article/pii/S0022283698920253**.

74 M. V. Shapovalov and R. L. Dunbrack, A smoothed backbone-dependent rotamer library for proteins derived from adaptive kernel density estimates and regressions, *Structure*, 2011, **19**(6), 844–858.

75 D. T. Keough, D. Hocková, A. Holý, L. M. J. Naesens, T. S. Skinner-Adams, J. D. Jersey, *et al.*, Inhibition of Hypoxanthine-Guanine Phosphoribosyltransferase by Acyclic Nucleoside Phosphonates: A New Class of Antimalarial Therapeutics, *J. Med. Chem.*, 2009, **52**(14), 4391–4399, DOI: **10.1021/jm900267n**.