Supplementary information for DeepRLI: a multi-objective framework for universal protein–ligand interaction prediction

Haoyu Lin,[†,§] Jintao Zhu,[†,§] Shiwei Wang,[‡] Yibo Li,[¶] Jianfeng Pei,[*,†] and Luhua Lai[*,†,‡,¶]

†*Center for Quantitative Biology, Academy for Advanced Interdisciplinary Studies, Peking University, Beijing 100871, China*

‡*Peking University Chengdu Academy for Advanced Interdisciplinary Biotechnologies, Chengdu, Sichuan 610213, China*

¶*BNLMS, State Key Laboratory for Structural Chemistry of Unstable & Stable Species, College of Chemistry and Molecular Engineering, Peking University, Beijing 100871, China*

§*These authors contributed equally to this work.*

E-mail: jfpei@pku.edu.cn; lhlai@pku.edu.cn

1 Supplementary overview of DeepRLI

The content of this section is a supplement to the main overview of DeepRLI section.

1.1 Ablation study on the cosine envelope

We train a version of the model without the cosine envelope using the same hyperparameter configuration. This version achieves a Pearson correlation coefficient of 0.838 on CASF-2016,

slightly lower than the performance with the cosine envelope ($R_{\mathrm{p}} = 0.849$). This indicates that the cosine envelope enhances the model's capability. Theoretically, its impact is more pronounced when training data is limited, as it helps prevent the model from converging to incorrect local minima of the loss function, which could otherwise result in assigning higher attention weights to distant nodes. When ample training data is available, the influence of the cosine envelope diminishes, as sufficient data enables the model to learn the correct relationships between interactions and distances.

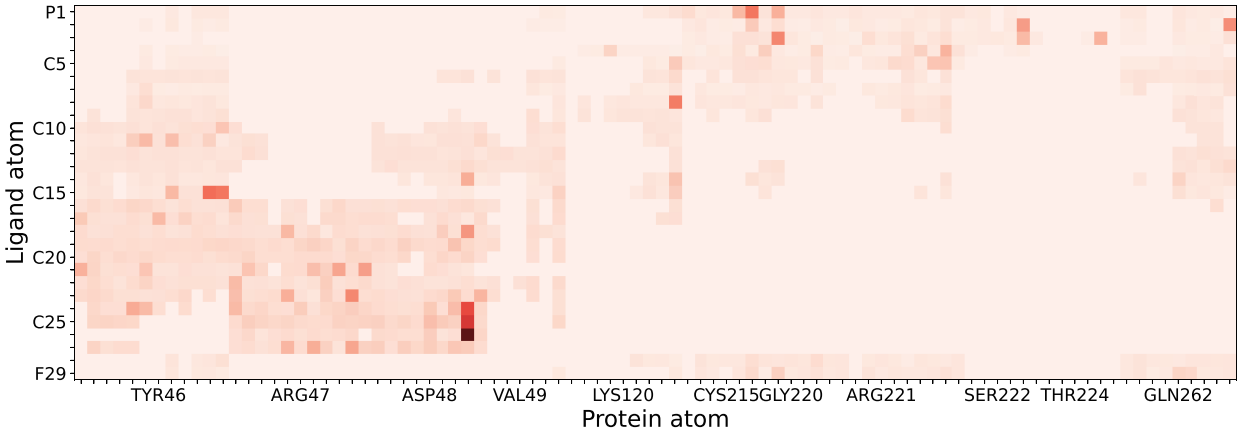

Figure S1: Visualization of interactions based on the attention weights from the final graph transformer layer of DeepRLI without cosine envelope. The heatmap illustrating [ligand atom]–[residue atom] interaction connections. Darker colors represent higher attention weights and more important interactions. For clarity, only part of interaction connections with a larger weight are shown.

The cosine envelope is integrated into the network after the softmax layer in the attention mechanism, so the sum of the weights applied to node embeddings is not constrained to equal to 1. Since the cosine envelope function is always less than or equal to 1, the total weights are typically less than 1. This feature ensures that when a central atom does not have strong interactions with surrounding atoms, these atoms are not assigned high weights. Beyond performance enhancement, the cosine envelope also improves model interpretability. As illustrated in Figure S1, compared to that illustrated in Figure 7d of the main text, the version of DeepRLI with the cosine envelope exhibits more focused attention weights, highlighting strong interactions and diminishing weaker ones.

2 Supplementary results and discussions

The content of this section is a supplement to the main results and discussions section.

2.1 Assessment of the model performance

2.1.1 Introduction to benchmarks

The performance of DeepRLI is tested on multiple recognized benchmarks, including benchmarks such as CASF-2016,[1] CSAR-NRC HiQ,[2] Merck FEP[3] and LIT-PCBA.[4]

CASF-2016 benchmark. The Comparative Assessment of Scoring Functions (CASF) benchmark was created by Cheng *et al.* and first published in 2009 as CASF-2007.[5] It has since been maintained and updated, with subsequent releases of CASF-2013[6,7] and the latest CASF-2016 version. In a nutshell, this dataset consists of different protein–ligand complexes with high-resolution crystal structures and reliable binding constants, obtained through systematically non-redundant sampling from the PDBbind database. And it is used to evaluate the performance of scoring functions regarding protein–ligand binding in the four previously mentioned aspects. Specifically, the CASF-2016 benchmark, which is the focus of this study, comprises an array of 285 high-quality protein–ligand complex crystal structures accompanied by reliable binding constants. Notably, these structures cover 57 different targets, each with five structures bound to different ligands. The structure-activity data pairs can be used for scoring and ranking capability assessment. Moreover, to meet the requirements for docking and screening ability tests, CASF-2016 also includes protein–ligand binding poses (decoys) generated by molecular docking programs. For each protein–ligand complex, a decoy set containing up to 100 ligand binding poses is generated for docking ability assessment. Additionally, each protein is cross-docked with another 280 ligands to obtain a larger decoy set suitable for screening ability assessment.

CSAR-NRC HiQ benchmark. The Community Structure-Activity Resource (CSAR)-

National Research Council of Canada (NRC) High-Quality (HiQ) benchmark was proposed in 2010 by Dunbar and many other researchers. CSAR aims to collect data from industry and academia that can be used to improve docking and scoring computational methods. The CSAR-NRC HiQ benchmark primarily serves to evaluate the efficacy of various scoring algorithms. Originally, the dataset encompassed 343 distinct protein–ligand complexes, divided into two sets: set1 with 176 and set2 with 167 complexes. Subsequently, in 2011, the dataset was expanded to include an additional set, set3, comprising 123 complexes, thereby augmenting the total count to 466 structure-activity datasets. A critical aspect to consider is the substantial overlap of complex structures within the CSAR-NRC HiQ dataset and our training set, PDBbindGS_HiQ. To mitigate the risk of artificially inflating the scoring performance of our model due to this redundancy, we have elected to exclude these overlapping complexes. This adjustment results in a revised dataset composition, with set1, set2, and set3 now containing 50, 36, and 75 complexes, respectively. Furthermore, this benchmark is often not evaluated in its complete state in other published works, for example, parts overlapping with the entire PDBbind general set are removed.[8] For ease of comparison, we also evaluate it on the same datasets as those works.

Merck FEP benchmark. Accurately ranking small molecules with subtle differences in binding efficacy to specific proteins plays an important role in the hit-to-lead and lead optimization stages of drug discovery. To address this challenging task, rigorous free energy simulation methods such as free energy perturbation (FEP), thermodynamic integration (TI), and λ-dynamics are employed for this purpose,[9] among which Schrödinger FEP+[10] is currently recognized as a mature and reliable program. In 2020, Merck KGaA published a benchmark for the assessment of relative free energy calculations, namely the Merck FEP benchmark, and tested the effects of large-scale prospective applications of FEP+.[3] This dataset collected a total of 8 pharmaceutical targets and 264 active ligands, with ligands belonging to a specific target having analogous skeletons and nuanced structural variations. Therefore, it can be used to further evaluate the ranking power of our model, especially the

possibility of its application in hit-to-lead and lead optimization.

LIT-PCBA benchmark. The LIT-PCBA is a recent benchmark specifically designed for the comparative evaluation of virtual screening. Compared to past analogous test sets, it relies on experimental biological activity data from the PubChem BioAssay database to support its crafting, thereby minimizing the presence of false positive and false negative compound data, which is common in past benchmarks due to insufficient experimental data and random selection of decoys. Moreover, it maintains a similar range and distribution of molecular properties for both active and inactive compounds, avoiding inappropriate evaluations of virtual screening methods due to inherent chemical biases. Therefore, the LIT-PCBA benchmark is currently a suitable dataset for evaluating the screening power of machine learning-based scoring functions (MLBSFs). Notably, it includes 15 targets, 7955 active compounds, and 2644022 inactive compounds. Such a hit rate (the ratio of active to inactive compounds) accurately reflects the real-world virtual screening scenario, greatly aiding in our further understanding of DeepRLI's screening effectiveness in challenging tasks. Since our model is based on 3D complex structures and currently lacks the capability for conformational sampling, in our tests, we first employ Glide SP[11,12] to ascertain various binding poses of small molecules with proteins, and then screen molecules guided by the predictions of DeepRLI.

0 Ligand Bias benchmark. The 0 Ligand Bias benchmark[13] is designed to evaluate whether MLBSFs rely on ligand-specific biases rather than generalizable biophysics by eliminating ligand-dependent information. It consists of 365 protein–ligand complexes where identical ligands (matched by InChI-Key) bind to different proteins. To ensure the benchmark challenges model reliance on ligand identity, clusters of identical ligands are selected with mean pK values between 6 and 7 (normalized across the dataset) and a variance in pK values greater than 1, removing cases where ligands have highly similar binding affinities. This setup forces models to disregard ligand-specific memorization and instead leverage protein features or interactions, testing their ability to generalize beyond ligand bias. By

excluding ligand-dependent cues, the benchmark rigorously assesses whether MLBSFs can capture meaningful protein–ligand interactions rather than dataset artifacts.

2.1.2 Evaluation on CASF-2016 benchmark

Figure S2 shows the results of the scoring, ranking, docking, and screening power of the DeepRLI model evaluated on the CASF-2016 benchmark.[1]

2.1.3 Evaluation on CSAR-NRC HiQ Benchmark

Figure S3 shows the results of the scoring power of the DeepRLI model evaluated on the CSAR-NRC HiQ benchmark.[2] The CSAR-NRC HiQ benchmark includes three protein–ligand complex structure datasets ("set1", "set2" and "set3"). We calculate not only the results for these three individual datasets, but also for their union ("sett"). To avoid interference from data leakage and facilitate comparison with other methods, we also evaluate the results on datasets with some complex structure data removed (Figures S4, S5 and S6).

2.1.4 Evaluation on Merck FEP benchmark

Table S1 shows the results of the ranking power of the DeepRLI model evaluated on the Merck FEP benchmark.[3]

2.2 Interpretation

2.2.1 PLIP results of 1BZC case

To validate the reasonableness of the attention mechanism-based interpretability of DeepRLI, we use the traditional rule-based protein–ligand interaction analysis program PLIP[15,16] to infer the potential key molecular interactions in the 1BZC complex example. Figure S7 presents the three-dimensional structure around the binding site of the 1BZC complex and the key interactions detected by PLIP. Furthermore, the details of these interactions are exhaustively listed in Tables S2, S3, and S4.

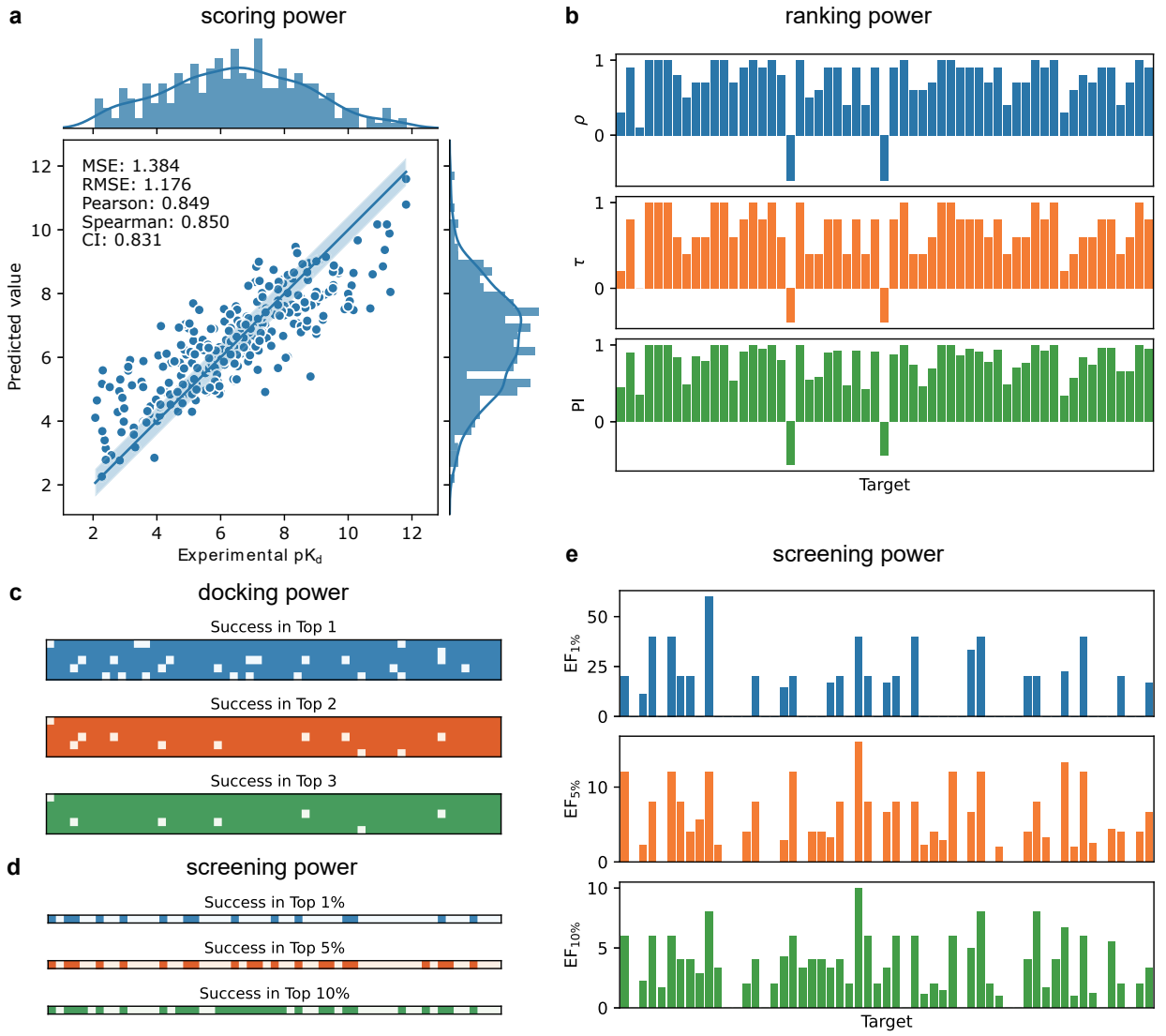

Figure S2: Performance of DeepRLI in scoring, ranking, docking, and screening on the CASF-2016 benchmark. (**a**) A correlation scatter plot depicting DeepRLI prediction of the experimental pK_d values for protein–ligand complexes. The light blue area surrounding the diagonal line represents the range of thermal fluctuations, specifically ±0.434, and data points falling within this range can be considered to be in good agreement with the experimental data. (**b**) Bar plots for three ranking metrics demonstrate DeepRLI's ability to rank active small molecules of various targets. The targets are arranged from left to right in alphabetical order of their PDB IDs. (**c**) Three heatmaps composed of 285 (5×57) small squares, where the squares from left to right and top to bottom correspond to complexes arranged in alphabetical order of their PDB IDs. Colored squares represent successfully docked complexes, i.e., those where one of the top n (1, 2, 3) poses predicted by DeepRLI have an RMSD less than 2 Å; in contrast, uncolored squares represent failures. (**d**) Similar to **c**, these heatmaps show whether active ligands are present in the top α (1%, 5%, 10%) small molecules predicted by DeepRLI. (**e**) Similar to **b**, the three bar graphs demonstrate DeepRLI's ability to enrich active small molecules in the top α (1%, 5%, 10%) across various targets.

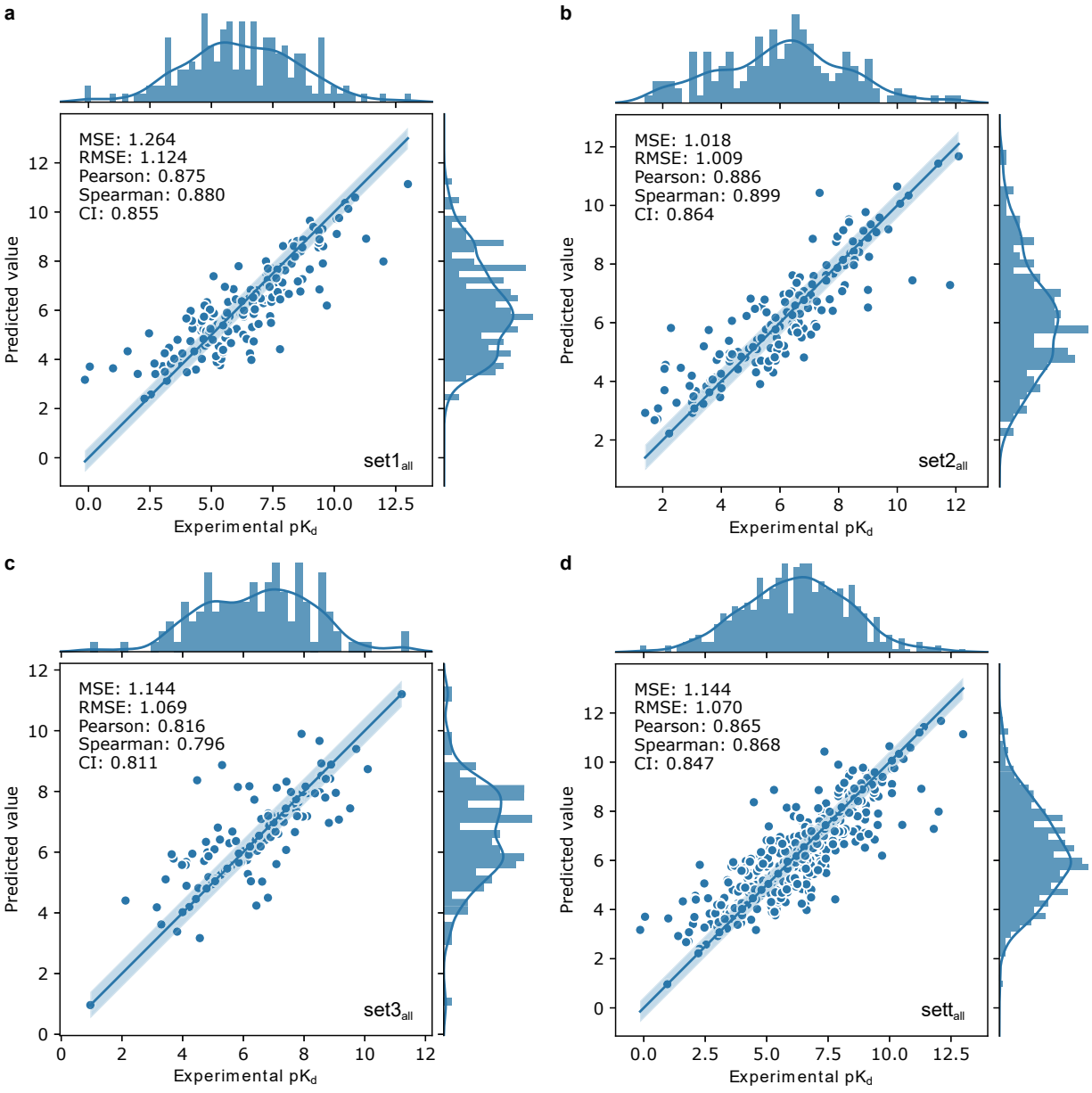

Figure S3: Scoring performance of DeepRLI on the CSAR-NRC HiQ benchmark. Scatter plots **a**–**d** show the correlation between predicted binding affinity values by DeepRLI and the actual experimental pK_d values for various types of protein–ligand complex structure datasets, which are sequentially from datasets "set1", "set2", "set3", and "sett", including all data of datasets (subscript "all").

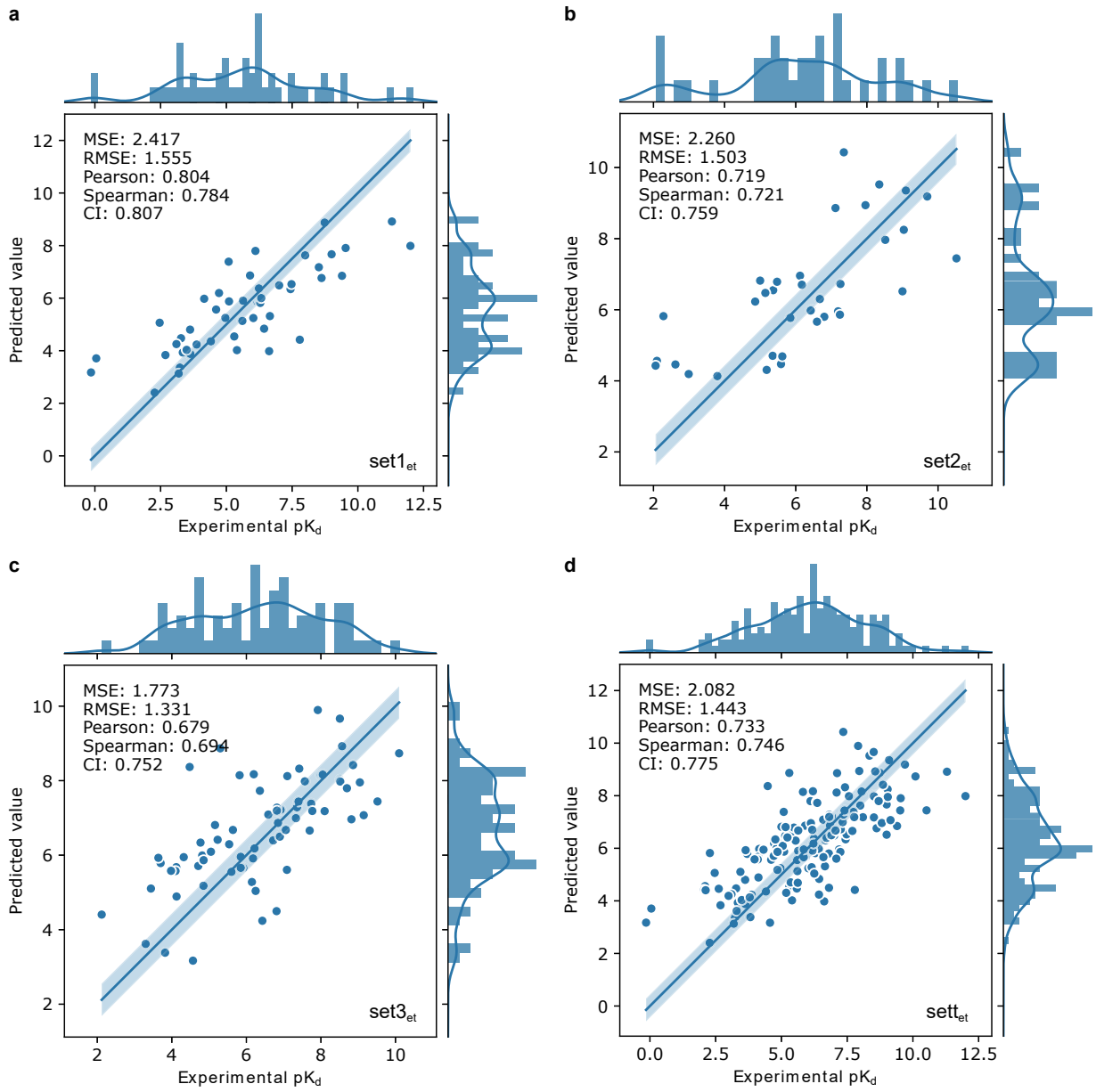

Figure S4: Scoring performance of DeepRLI on the CSAR-NRC HiQ benchmark. Scatter plots **a**–**d** show the correlation between predicted binding affinity values by DeepRLI and the actual experimental pK_d values for various types of protein–ligand complex structure datasets, which are sequentially from datasets "set1", "set2", "set3", and "sett", excluding training set part of datasets (subscript "et").

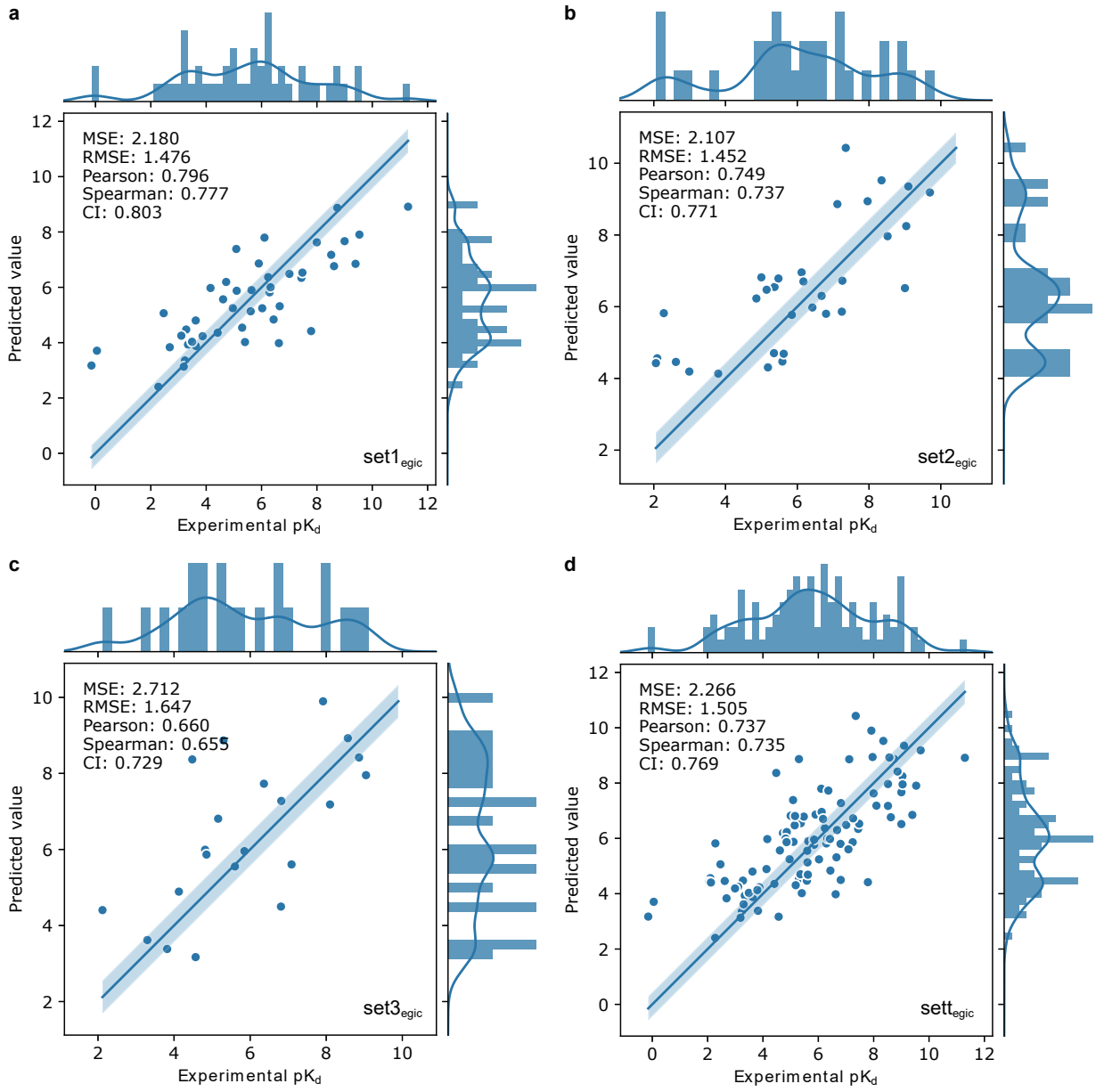

Figure S5: Scoring performance of DeepRLI on the CSAR-NRC HiQ benchmark. Scatter plots **a**–**d** show the correlation between predicted binding affinity values by DeepRLI and the actual experimental pK_d values for various types of protein–ligand complex structure datasets, which are sequentially from datasets "set1", "set2", "set3", and "sett", exclude general set part but include core set part of datasets (subscript "egic").

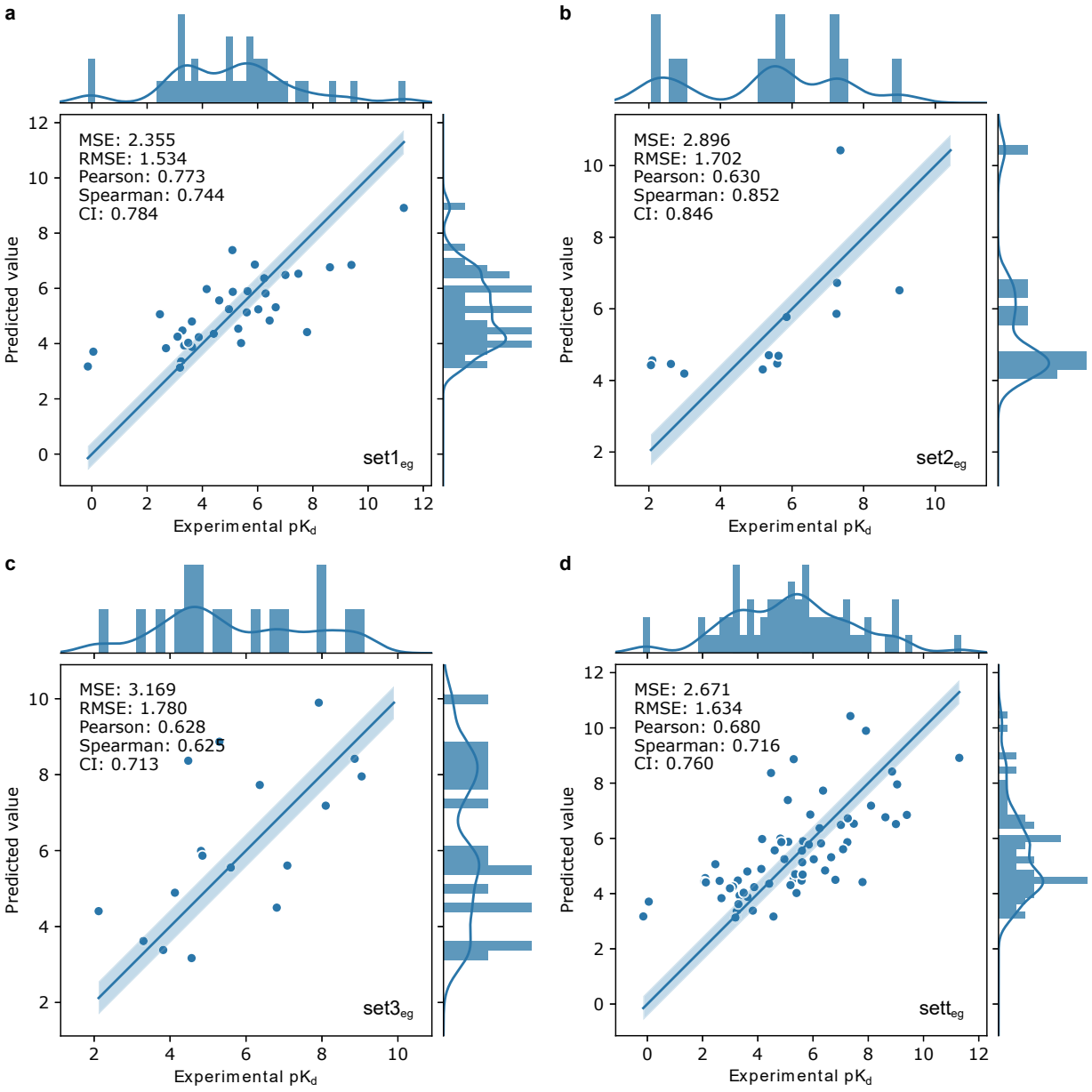

Figure S6: Scoring performance of DeepRLI on the CSAR-NRC HiQ benchmark. Scatter plots **a**–**d** show the correlation between predicted binding affinity values by DeepRLI and the actual experimental pK_d values for various types of protein–ligand complex structure datasets, which are sequentially from datasets "set1", "set2", "set3", and "sett", exclude general set part of datasets (subscript "eg").

Table S1: The ranking power, measured by Spearman correlation coefficient (ρ), of several representative scoring functions on the Merck FEP benchmark. Apart from PBCNet and DeepRLI, the data for all other models are from Shen *et al.*[8] PBCNet is a model recently developed specifically for the task of relative binding affinity prediction.[14] The best result in each column is highlighted in bold

Method	CDK8	c-Met	Eg5	HIF-2α	PFKFB3	SHP-2	SYK	TNKS2	mean
AutoDock4	0.629	0.324	−0.397	0.376	0.530	0.609	0.544	0.558	0.397
Vina	**0.849**	−0.257	−0.520	**0.493**	0.546	0.569	0.519	0.538	0.342
Vinardo	0.782	−0.359	−0.475	0.371	0.515	0.490	0.379	0.305	0.251
Glide SP	0.345	0.378	−0.111	0.445	0.480	0.542	−0.006	0.316	0.299
Glide XP	0.617	0.165	0.017	0.410	0.513	0.490	0.124	0.582	0.365
X-Score	0.406	0.531	−0.316	0.224	0.430	−0.030	**0.689**	**0.669**	0.325
MM-GBSA	0.649	0.499	−0.002	0.282	0.554	0.585	0.108	0.158	0.354
$\Delta_{\text{Lin_F9}}$XGB	0.826	0.077	−0.099	0.480	**0.603**	**0.640**	0.103	0.458	0.386
Pafnucy	0.406	0.531	−0.316	0.224	0.430	−0.030	**0.689**	0.669	0.325
GenScore	0.675	0.677	0.275	0.437	0.571	0.338	0.144	0.578	0.462
PBCNet	0.63	**0.76**	**0.58**	0.30	0.47	0.56	0.48	0.32	**0.51**
DeepRLI	0.513	0.745	−0.024	0.459	0.577	0.639	0.441	0.331	0.460

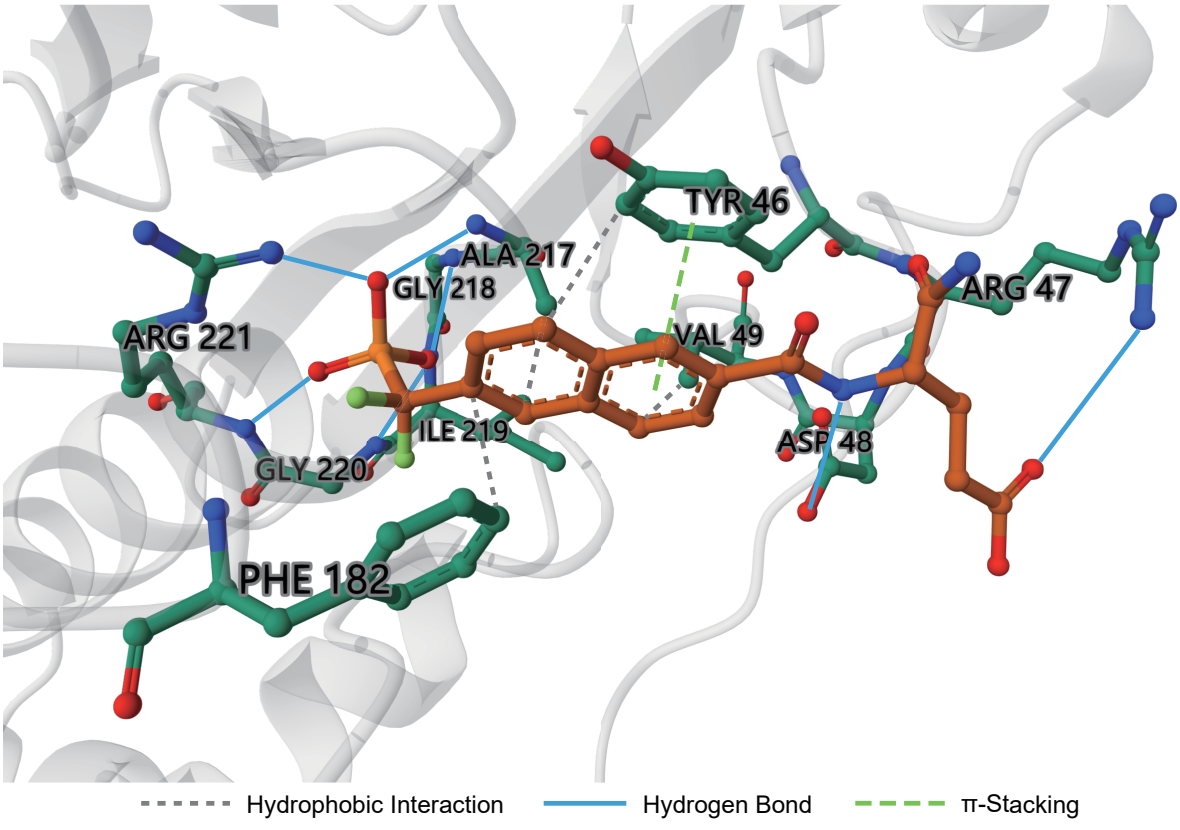

Figure S7: The 3D molecular structure surrounding the binding site of the 1BZC complex, as well as the potential intermolecular non-covalent interactions present within the complex detected by the the PLIP non-covalent interaction analysis tool.

Table S2: The potential intermolecular hydrophobic interactions present within the 1BZC complex identified through the utilization of the PLIP non-covalent interaction analysis tool[a]

Residue	Distance (Å)	Ligand atom	Protein atom
TYR46	3.94	2438	377
VAL49	3.91	2442	405
PHE182	3.35	2436	1527
ALA217	3.45	2435	1786

[a] In the PLIP analysis results, the atoms are represented by the numbering in the PDB file.

Table S3: The potential intermolecular hydrogen bonds present within the 1BZC complex identified through the utilization of the PLIP non-covalent interaction analysis tool[b]

Residue	Distance H-A (Å)	Distance D-A (Å)	Donor angle (ř)	Donor atom	Acceptor atom
ARG47	3.27	4.03	135.35	389 [Ng+]	2456 [O.co2]
ARG47	2.27	3.19	155.66	380 [Nam]	2451 [O2]
ASP48	2.94	3.90	165.69	2447 [Nam]	398 [O3]
ALA217	1.87	2.81	159.54	1782 [Nam]	2431 [O3]
GLY218	2.80	3.30	111.55	1787 [Nam]	2429 [O2]
ILE219	2.17	3.10	156.43	1791 [Nam]	2429 [O2]
GLY220	1.96	2.94	172.50	1799 [Nam]	2429 [O2]
ARG221	1.84	2.81	169.06	1803 [Nam]	2430 [O3]
ARG221	1.79	2.72	155.79	1813 [Ng+]	2431 [O3]

[b] The atom serial numbers in the PLIP analysis results are accompanied by brackets, within which are recorded the atom types according to SYBYL[17] or IDATM.[18]

Table S4: The potential intermolecular π-stackings present within the 1BZC complex identified through the utilization of the PLIP non-covalent interaction analysis tool

Residue	Distance (Å)	Angle (°)	Offset (Å)	Stacking type	Ligand atoms
TYR46	3.92	21.79	0.63	Parallel	2439, 2440, 2441, 2442, 2443, 2444

2.2.2 More examples for interpretation

Case: 2FVD

The 2FVD system represents cyclin-dependent kinase 2 (CDK2) complexed with inhibitor [4-amino-2-(1-methanesulfonylpiperidin-4-ylamino)pyrimidin-5-yl](2,3-difluoro-6-methoxyphenyl)methanone (LIA).[19] CDKs are serine/threonine kinases regulating cell cycle progression.[20] Their dysregulation in cancers makes CDK inhibition a promising therapeutic strategy.[21]

Table S5: PLIP-detected hydrophobic interactions in 2FVD complex[a]

Residue	Distance (Å)	Ligand Atom	Protein Atom
VAL18	3.99	2312	141
ALA31	3.74	2304	242
PHE82	3.92	2296	570
LEU134	3.80	2304	990
LEU134	3.78	2298	991
ASP145	3.44	2313	1067

[a]Atom numbering follows PDB file conventions.

Table S6: PLIP-detected hydrogen bonds in 2FVD complex[b]

Residue	H-A Distance (Å)	D-A Distance (Å)	Donor Angle (ř)	Donor Atom	Acceptor Atom
GLU81	2.16	2.74	114.11	2319 [N3]	554 [O2]
LEU83	2.70	3.68	171.25	571 [Nam]	2306 [N2]
LEU83	2.10	2.76	122.80	2300 [Npl]	574 [O2]
ASP86	2.15	3.05	151.92	598 [Nam]	2322 [O2]
LYS89	2.77	3.21	106.67	631 [N3]	2323 [O2]

[b]Atom types in brackets follow SYBYL/IDATM nomenclature.

Table S7: PLIP-detected salt bridges in 2FVD complex

Residue	Distance (Å)	Ligand Group	Ligand Atoms
ASP86	4.92	Tertamine	2294

Figure S8 demonstrates DeepRLI's interpretability analysis for 2FVD. Residues ILE10, VAL18, LYS33, PHE80, PHE82, HIS84, GLN85, ASP86, LYS89, and ASP145 show high

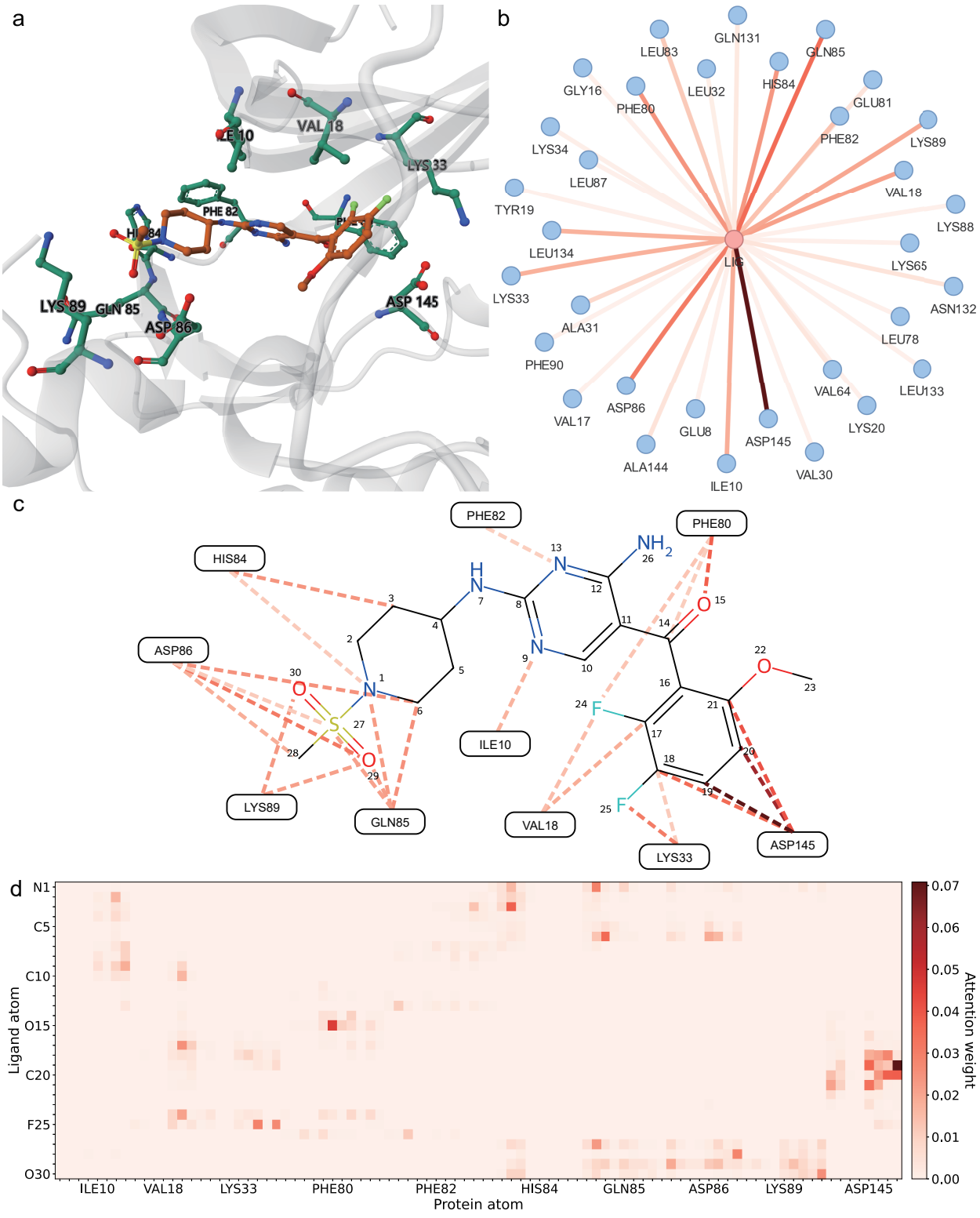

Figure S8: Visualization of interactions based on attention weights from DeepRLI's final graph transformer layer. Color intensity reflects attention weight magnitude, with darker colors indicating higher weights. Shown are results for PDB ID 2FVD complex. **a**, 3D binding site structure with high-attention residues in ball-and-stick representation. **b**, Circular layout of residue-level interaction importance. **c**, Ligand-centric view of residue-atom interactions. **d**, Atom-level interaction heatmap. Only significant interactions are displayed in (c) and (d) for clarity.

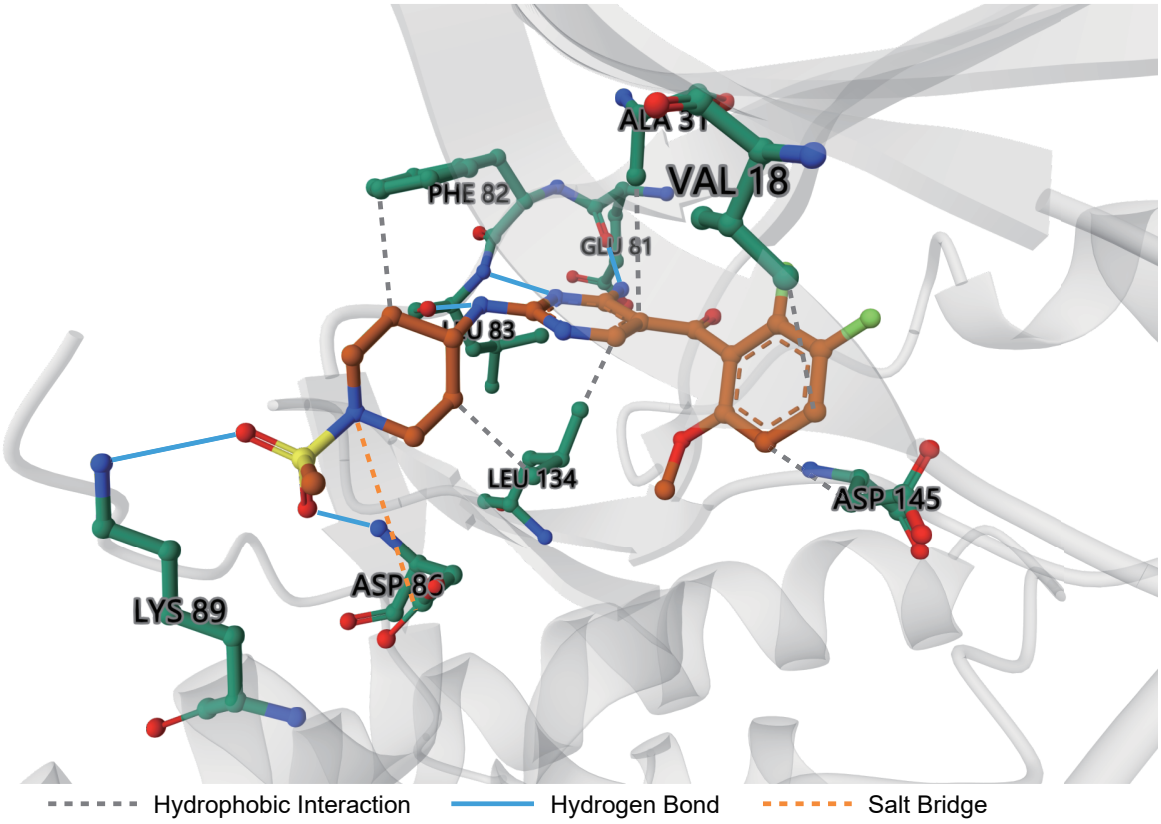

Figure S9: 3D molecular structure of 2FVD binding site with PLIP-detected non-covalent interactions.

attention weights, indicating critical contributions to binding affinity prediction. Atomic-level analysis reveals strong attention between VAL18/ASP145 and LIA's 2,3-difluoro-6-methoxyphenyl group, and between PHE82/ASP86/LYS89 and LIA's 1-methanesulfonylpiperidin-4-ylamino moiety. PLIP analysis (Tables S5-S7) confirms hydrophobic interactions with VAL18/PHE82/ASP145, hydrogen bonds/salt bridges with ASP86, and hydrogen bonds with LYS89, aligning with DeepRLI's attention patterns.

Case: 3ARP

The 3ARP system contains *Vibrio harveyi* chitinase A complexed with dequalinium (DEQ).[22] Chitinases catalyze β-(1,4)-linked N-acetylglucosamine polymer degradation, crucial for fungal cell wall maintenance.[22]

Table S8: PLIP-detected hydrophobic interactions in 3ARP complex

Residue	Distance (Å)	Ligand Atom	Protein Atom
1TRP68	3.64	4547	1159
TRP168	3.83	4546	1161
TRP168	3.62	4544	1162
TRP168	3.55	4572	1160
VAL205	3.80	4551	1464
TRP275	3.72	4568	2021
TRP275	3.88	4567	2022
TRP275	3.96	4575	2029
TRP275	3.56	4563	2028
THR276	3.77	4545	2037
LEU277	3.91	4546	2044
ASP392	3.74	4575	2984
TRP397	3.75	4558	3030
TRP397	3.59	4554	3026
TRP397	3.30	4560	3023
TRP570	3.76	4570	4351

DeepRLI analysis (Fig. S10) highlights TRP168, VAL205, HIS228, TRP275, THR276, LEU277, ASP392, and TRP397 as key contributors. High attention weights focus on interactions between TRP168/TRP275/TRP397 and DEQ's 4-amino-2-methylquinolinium groups. PLIP results (Tables S8-S10) confirm extensive hydrophobic contacts, water bridges, and

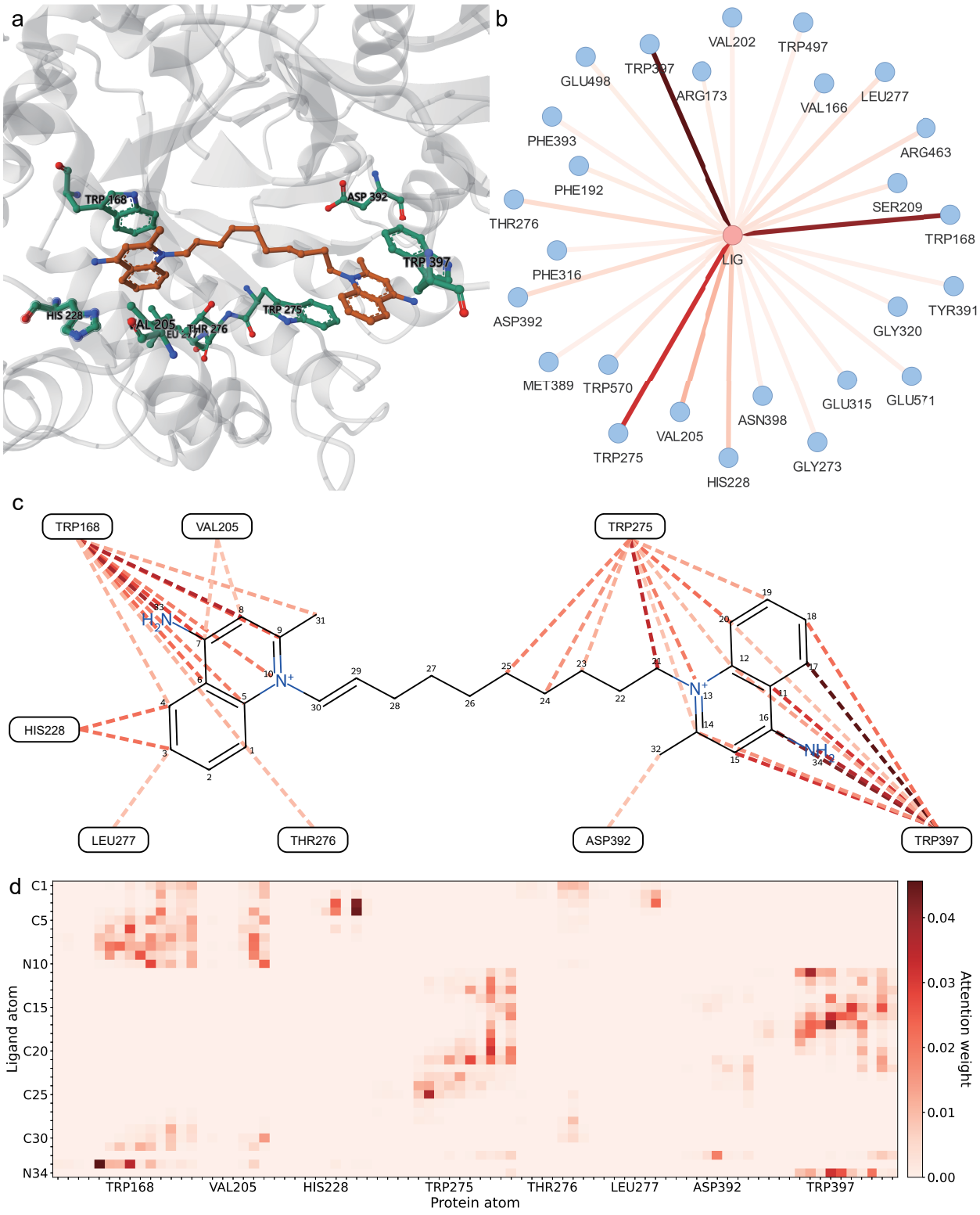

Figure S10: Attention weight visualization for 3ARP complex. **a**, Binding site 3D structure with high-attention residues. **b**, Residue-level interaction network. **c**, Ligand-atom interaction mapping. **d**, Atomic interaction heatmap.

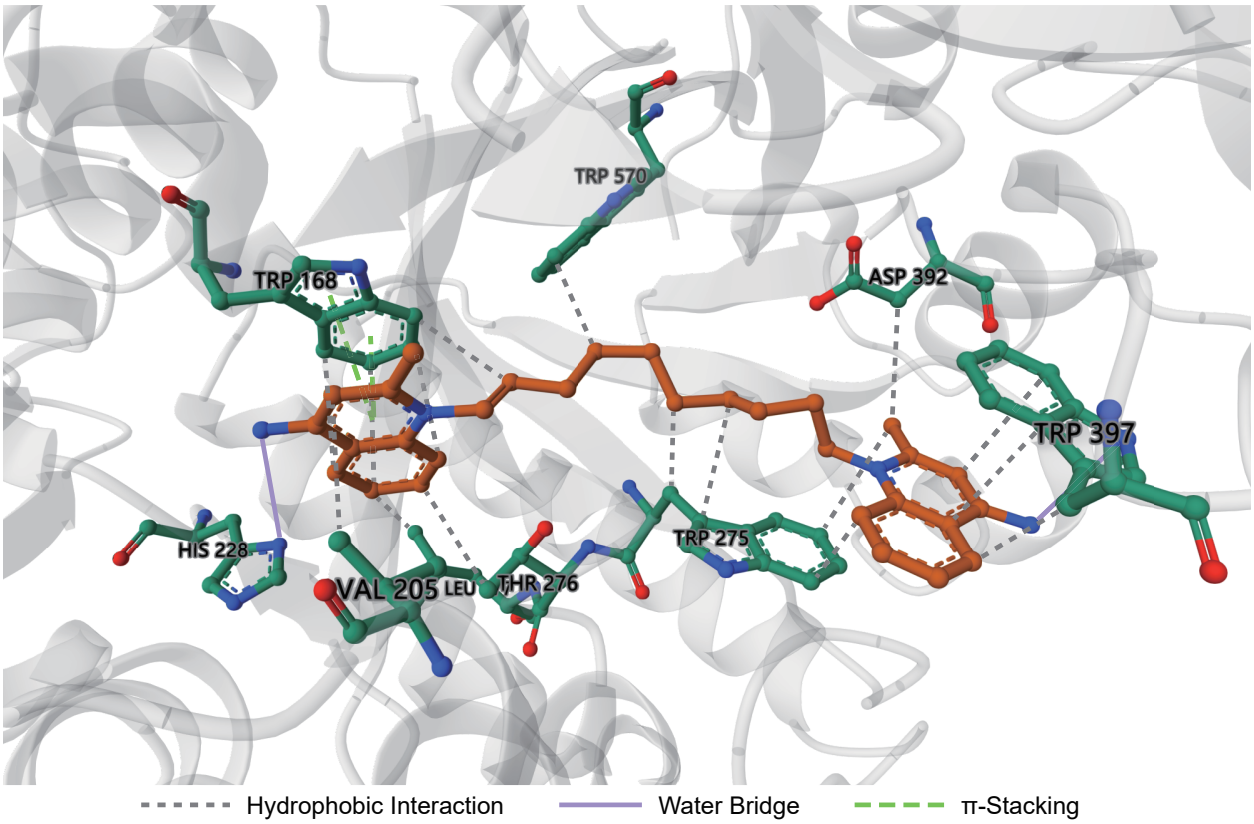

Figure S11: PLIP-detected interactions in 3ARP binding site.

Table S9: PLIP-detected water bridges in 3ARP complex

Residue	A-W (Å)	D-W (Å)	Donor Angle (ř)	Water Angle (ř)	Donor Atom	Acceptor Atom	Water Atom
HIS228	4.04	2.84	148.49	75.07	4576 [Npl]	1642 [N2]	4900
TRP397	4.02	3.17	168.97	123.09	4577 [Npl]	3027 [N2]	5117

Table S10: PLIP-detected π-stacking in 3ARP complex

Residue	Distance (Å)	Angle (ř)	Offset (Å)	Type	Ligand Atoms
TRP168	3.78	0.75	1.56	Parallel	4548-4553
TRP168	3.52	1.72	0.70	Parallel	4548-4553

π-stacking interactions matching DeepRLI's attention patterns.

Case: 4DE2

The 4DE2 system contains CTX-M-9 class A β-lactamase complexed with inhibitor 3-[(dimethylamino)methyl]-N-[3-(1H-tetrazol-5-yl)phenyl]benzamide (DN3).[23] β-lactamases mediate bacterial resistance to β-lactam antibiotics by catalyzing their hydrolysis, making β-lactamase inhibitors crucial for combating antimicrobial resistance.[23]

Table S11: PLIP-detected hydrophobic interactions in 4DE2 complex[a]

Residue	Distance (Å)	Ligand Atom	Protein Atom
TYR105	3.99	4044	606
TYR105	3.57	4043	608
PRO167	3.98	4037	1074
ASN170	3.67	4037	1096

[a]Atom numbering follows PDB file conventions.

Table S12: PLIP-detected hydrogen bonds in 4DE2 complex[b]

Residue	H-A Distance (Å)	D-A Distance (Å)	Donor Angle (ř)	Donor Atom	Acceptor Atom
SER70	3.07	3.78	131.72	329 [O3]	4049 [Nar]
ASN104	1.93	2.91	172.34	601 [Nam]	4026 [O2]
SER130	2.67	3.62	167.46	803 [O3]	4048 [Nar]
ASN170	2.82	3.25	107.49	1099 [Nam]	4026 [O2]
LYS234	3.32	4.02	127.17	1571 [N3]	4049 [Nar]
GLY236	3.35	3.82	111.34	4047 [Nar]	1582 [O2]
SER237	3.49	3.86	105.08	4038 [Nam]	1586 [O2]

[b]Atom types in brackets follow SYBYL/IDATM nomenclature.

Table S13: PLIP-detected water bridges in 4DE2 complex

Residue	A-W (Å)	D-W (Å)	Donor Angle (ř)	Water Angle (ř)	Donor Atom	Acceptor Atom	Water Atom
THR235	3.82	4.00	105.00	103.76	1577 [O3]	4046 [Nar]	4148
THR235	3.73	4.00	105.00	136.91	1577 [O3]	4048 [Nar]	4148
SER237	3.92	2.87	164.97	101.13	1583 [Nam]	4049 [Nar]	4305
SER237	3.69	2.87	164.97	127.76	1583 [Nam]	4046 [Nar]	4305

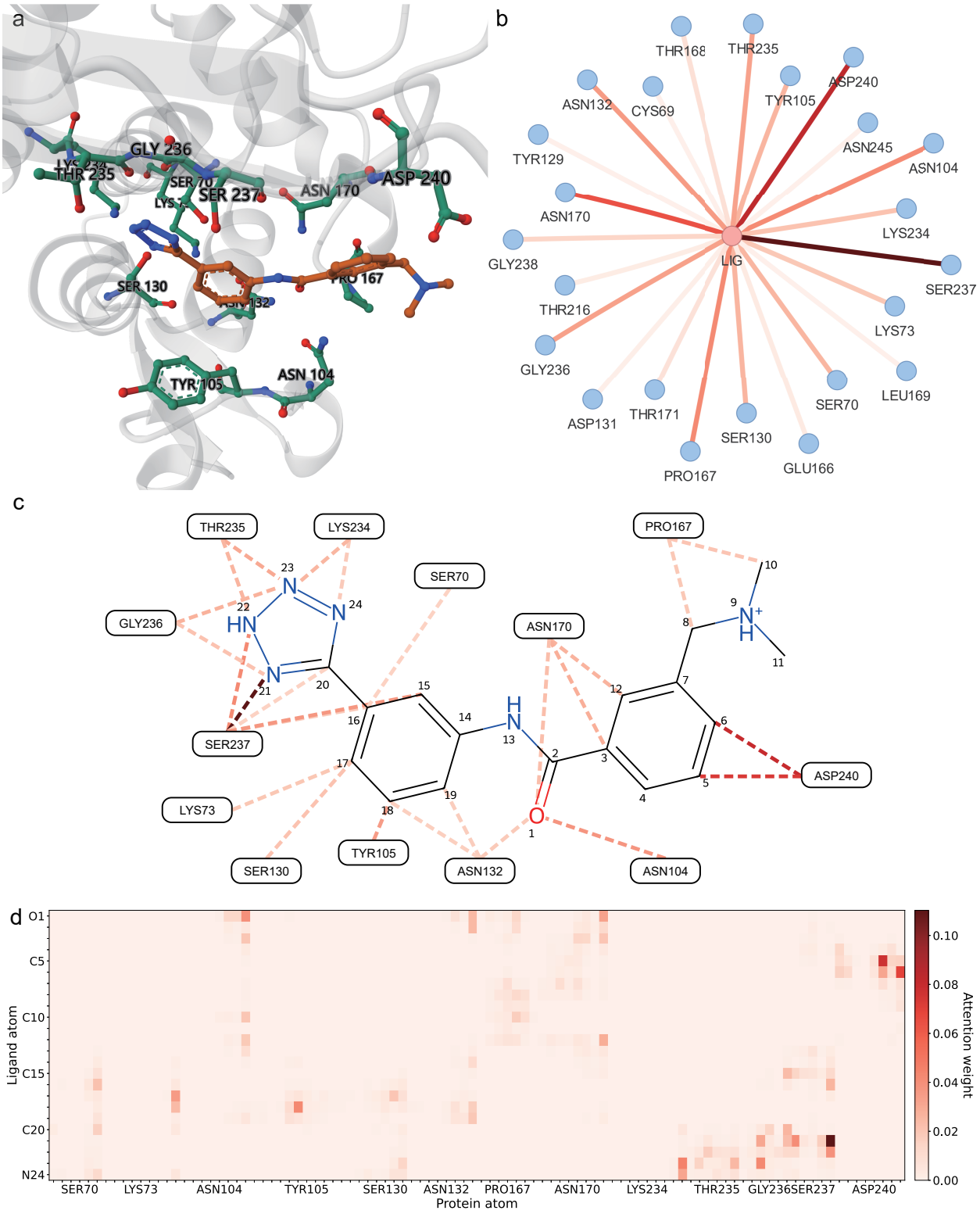

Figure S12: Attention weight visualization for 4DE2 complex. **a**, 3D binding site structure with high-attention residues in ball-and-stick representation. **b**, Circular layout of residue-level interaction importance. **c**, Ligand-centric view of residue-atom interactions. **d**, Atomic interaction heatmap (partial display for clarity).

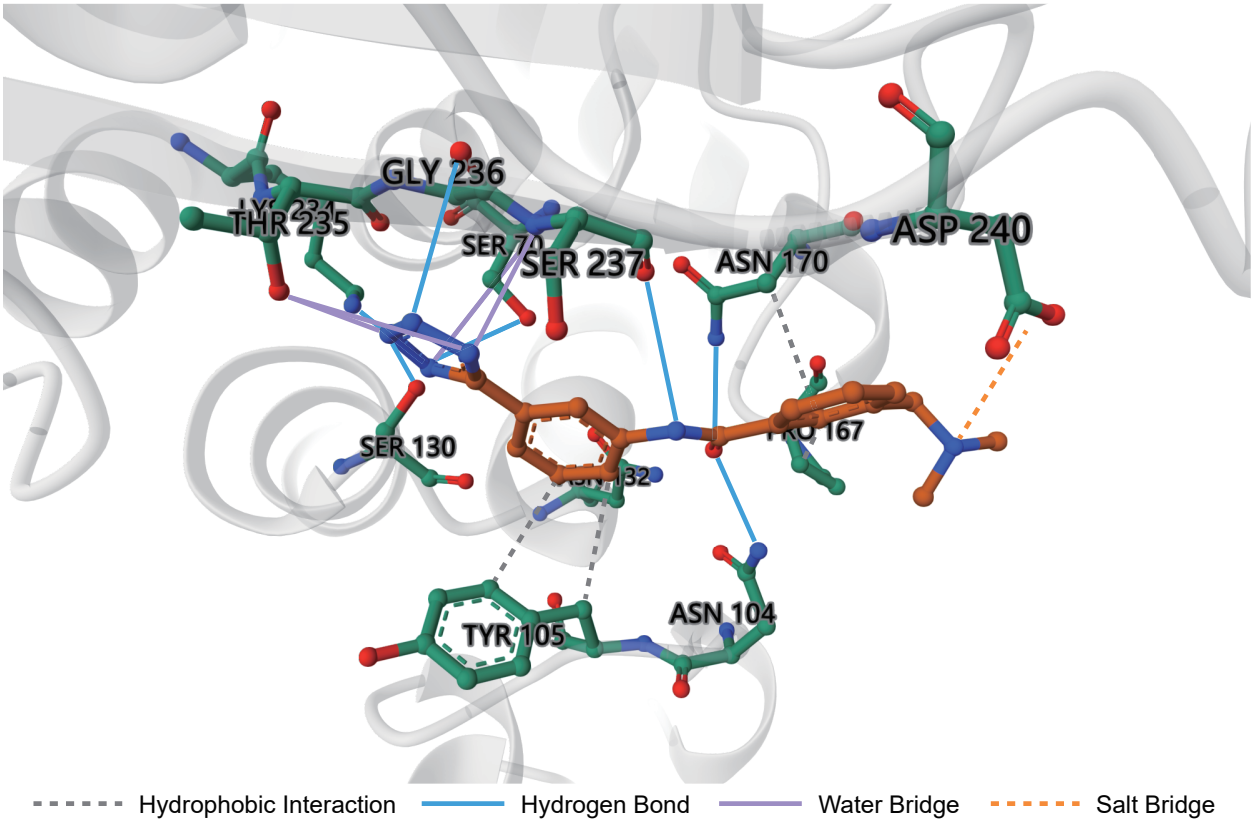

Figure S13: PLIP-detected interactions in 4DE2 binding site.

Table S14: PLIP-detected salt bridges in 4DE2 complex

Residue	Distance (Å)	Ligand Group	Ligand Atoms
ASP240	5.11	Tertamine	4034

Figure S12 reveals DeepRLI's attention patterns for 4DE2 complex. Key residues SER70, LYS73, ASN104, TYR105, SER130, ASN132, PRO167, ASN170, LYS234, THR235, GLY236, SER237, and ASP240 demonstrate high attention weights. Atomic-level analysis shows strong focus on interactions between SER70/TYR105/SER130/LYS234/THR235/GLY236 /SER237 and DN3's 3-(1H-tetrazol-5-yl)phenyl group, and between PRO167/ASN170/AS P240 and DN3's dimethylaminomethyl benzamide moiety. PLIP analysis (Tables S11-S14) confirms hydrophobic interactions with TYR105/PRO167/ASN170, hydrogen bonds with A SN104/SER130/ASN170/GLY236/SER237, water bridges with THR235/SER237, and salt bridges with ASP240, aligning precisely with DeepRLI's attention allocation.

Case: 4WIV

The 4WIV system represents human BRD4 bromodomain complexed with inhibitor N-tert-butyl-2-[4-(3,5-dimethyl-1,2-oxazol-4-yl)phenyl]imidazo[1,2-a]pyrazin-3-amine (3P2).[24] As a BET family member, BRD4 functions as transcriptional coactivator mediating signal transduction from master regulators (e.g., MYC in cancer, NFB in inflammation) to RNA Pol II.[25] Dysregulated BET bromodomain activity contributes to malignancies, making its inhibition a promising therapeutic strategy.[24]

Table S15: PLIP-detected hydrophobic interactions in 4WIV complex[a]

Residue	Distance (Å)	Ligand Atom	Protein Atom
PRO82	3.96	1075	341
PRO82	3.74	1064	340
PHE83	3.47	1053	351
VAL87	3.92	1051	385
LEU92	3.79	1064	421
LEU94	3.61	1054	437
TYR139	3.80	1054	824
ILE146	3.89	1053	877
ILE146	3.70	1061	875

[a]Atom numbering follows PDB file conventions.

DeepRLI's interpretability analysis (Fig. S14) identifies critical contributions from TRP81, PRO82, PHE83, VAL87, LEU92, CYS136, TYR139, ASN140, ASP145, ILE146, and MET149.

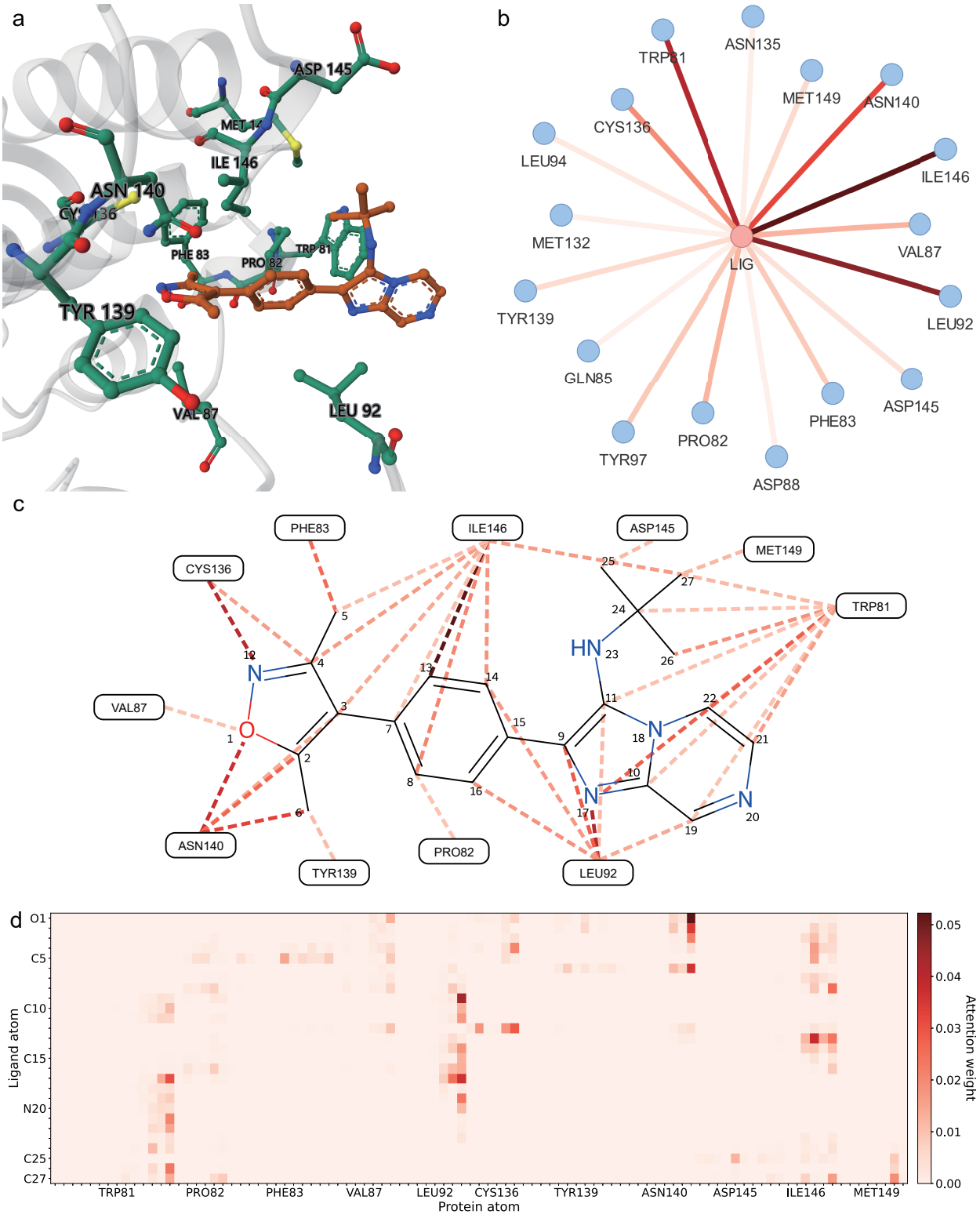

Figure S14: Interaction visualization through attention weights from DeepRLI's final graph transformer layer. Color intensity reflects weight magnitude, with darker hues indicating higher importance. Results shown for PDB ID 4WIV complex. **a**, Binding site 3D structure highlighting high-attention residues in ball-and-stick representation. **b**, Circular diagram of residue-level interaction importance. **c**, 2D ligand-centric view of residue-atom interactions. **d**, Atomic interaction heatmap. Panels (c) and (d) display only prominent interactions for clarity.

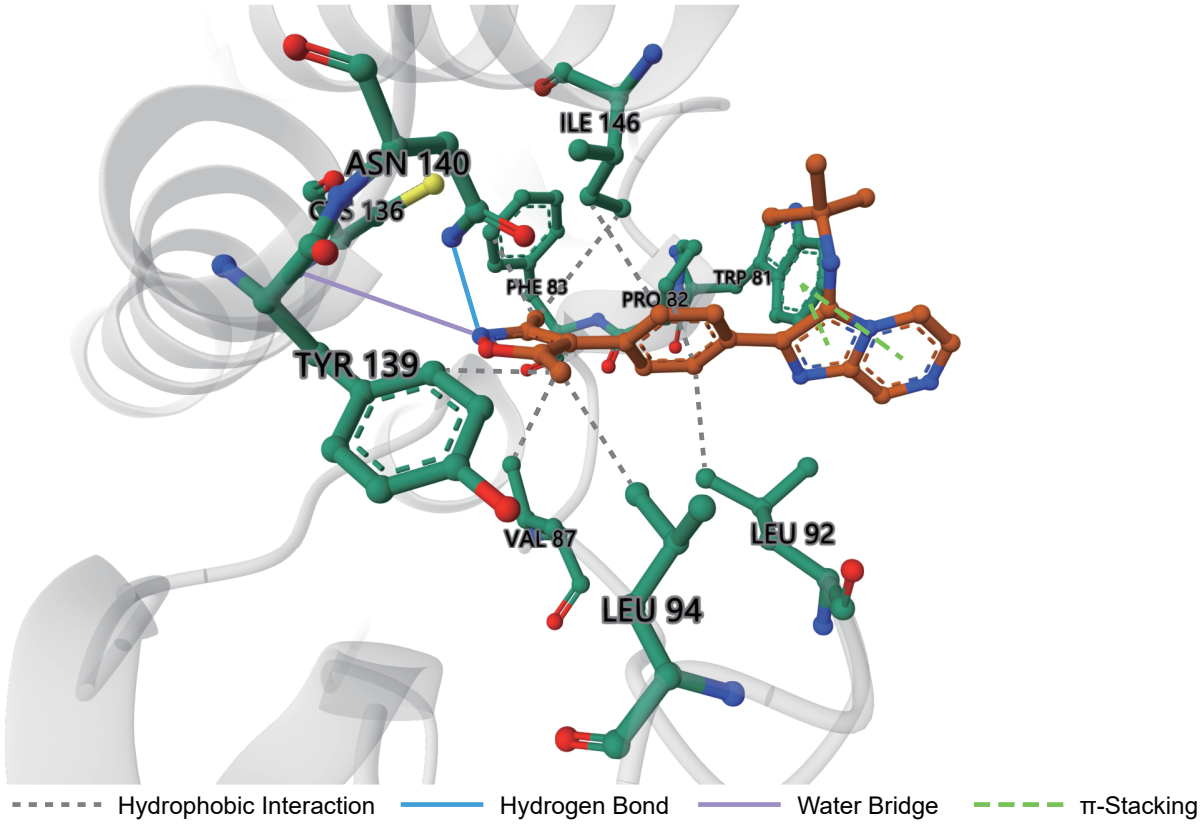

Figure S15: 3D molecular structure of 4WIV binding site with PLIP-detected non-covalent interactions.

Table S16: PLIP-detected hydrogen bonds in 4WIV complex[b]

Residue	H-A (Å)	D-A (Å)	Donor Angle (ř)	Donor Atom	Acceptor Atom
ASN140	3.00	3.84	143.91	836 [Nam]	1060 [Nar]

[b]Atom types in brackets follow SYBYL/IDATM nomenclature.

Table S17: PLIP-detected water bridges in 4WIV complex

Residue	A-W (Å)	D-W (Å)	Donor Angle (ř)	Water Angle (ř)	Donor Atom	Acceptor Atom	Water Atom
CYS136	3.89	3.71	104.05	88.37	791 [Nam]	1060 [Nar]	1183

Table S18: PLIP-detected π-stacking interactions in 4WIV complex

Residue	Distance (Å)	Angle (ř)	Offset (Å)	Type	Ligand Atoms
TRP81	4.86	83.53	1.08	T-shaped	1057-1066
TRP81	5.09	83.03	1.78	T-shaped	1058-1070

Atomic-level attention weights highlight: (1) Interactions between PHE83/VAL87/CYS136/TYR139/ASN140 and 3P2's 3,5-dimethyl-1,2-oxazol-4-yl group; (2) PRO82/LEU92/ILE146 interactions with the phenyl moiety; (3) TRP81/LEU92 interactions with imidazo[1,2-a]pyrazine portion. PLIP validation (Tables S15-S18) confirms: Hydrophobic contacts with PRO82/PHE83/VAL87/LEU92/TYR139/ILE146; Hydrogen bonding with ASN140; Water bridges with CYS136; π-stacking with TRP81. This alignment substantiates DeepRLI's capability to identify biologically significant interaction patterns for affinity prediction.

Through these five case studies, DeepRLI demonstrates interpretability through attention weight visualization from graph transformer layers. The significant interaction contributions identified by DeepRLI align remarkably with conventional PLIP analysis, validating the model's capability to capture critical protein–ligand interaction features. This correspondence confirms that graph representations learned through graph convolutional networks maintain strong correlation with binding affinities, enabling DeepRLI's robust performance across diverse prediction tasks.

3 Supplementary methods

The content of this section supplements the methods part of the main text.

3.1 Input features

Table S19 and Table S20 list the node features and edge features of the model input respectively.

3.2 Datasets

Table S21 lists the data used in the training process of DeepRLI.[26] The minimal input for training consists of a data unit made up of a crystal structure-activity pair, a re-docked positive structure, a re-docked negative structure, and a cross-docked negative structure.

Table S19: List of node features

Feature name	Feature vector
affiliation	0 or 1 (for protein or ligand)
symbol	C, N, O, F, P, S, Cl, Br, I, Met, Unk (one hot)
hybridization	s, sp, sp^2, sp^3, sp^3d, sp^3d^2 (one hot)
formal charge	-2, -1, 0, 1, 2, 3, 4 (one hot)
degree	0, 1, 2, 3, 4, 5 (one hot)
is a donor	0 or 1
is an acceptor	0 or 1
is negative ionizable	0 or 1
is positive ionizable	0 or 1
is Zn binder	0 or 1
is aromatic	0 or 1
is a hydrophobe	0 or 1
is a lumped hydrophobe	0 or 1

Table S20: List of edge features

Feature name	Feature vector
is intermolecular	0 or 1
is covalent	0 or 1
bond type	single, double, triple, aromatic (one hot)
distance	(gaussian smearing, 33)

In our actual training process, the data unit also includes an additional crystal structure-activity pair to further enhance the model's scoring performance.

Table S21: List of data units for training

No.	Native ID	Re-docked Positive IDs	Re-docked Negative IDs	Cross-docked Negative IDs
1	10gs	[10gs_27, 10gs_6]	[10gs_1, 10gs_10, 10gs_11, ...]	[10gs-1a9q, 10gs-1bn1, 10gs-1bnv, ...]
2	184l	[184l_1, 184l_12, 184l_13, ...]	[184l_11, 184l_14, 184l_15, ...]	[184l-1ghy, 184l-1t4v, 184l-3l0v, ...]
⋮	⋮	⋮	⋮	⋮
4156	966c	[966c_1, 966c_21]	[966c_11, 966c_12, 966c_13, ...]	[966c-1c1r, 966c-1d09, 966c-1d3d, ...]

+ Supplementary Native IDs (randomly select one for each row): [1afl, 1avn, 1bai, ...]

3.3 Derivation of contrastive loss

In order to be able to take advantage of the unknown exact pK_d data, we need to construct a loss function suitable for them. Here we explore the origin of the mean squared error (MSE) from the perspective of probability distribution, and then generalize this method to data with fuzzy true values.

In the ordinary affinity prediction model training, the dataset is a set of $\{(x_i, y_i^{\mathrm{true}})\}_{i=1}^N$. And our purpose is to build a neural network $y^{\mathrm{pred}} = \mu_\theta(x)$ so that the predicted value is as close as possible to the real value, that is, the loss function $\mathcal{L} = \frac{1}{N}\sum_{i=1}^N (y_i^{\mathrm{pred}} - y_i^{\mathrm{true}})^2$ is as small as possible. If we regard the data as independent and identically distributed random variables, that is

$$\{(x_i, y_i^{\mathrm{true}})\}_{i=1}^N \overset{i.i.d.}{\sim} p_\mathrm{d}(x, y), \tag{S1}$$

then the problem can be reformulated as constructing a neural network so that the predicted distribution $p_\theta(x, y)$ is as close as possible to the real distribution:

$$p_\theta(x, y) \to p_\mathrm{d}(x, y). \tag{S2}$$

That is, the optimization goal is to reduce the difference between these two distributions,

$$\mathcal{L} = D(p_{\mathrm{d}} \| p_\theta). \tag{S3}$$

And the difference can be quantified by the Kullback-Leibler (KL) divergence:

$$\begin{aligned} D_{\mathrm{KL}}(p_{\mathrm{d}} \| p_\theta) &= \iint p_{\mathrm{d}}(x,y) \ln \frac{p_{\mathrm{d}}(x,y)}{p_\theta(x,y)} \mathrm{d}x\mathrm{d}y & \text{(S4)} \\ &= -\iint p_{\mathrm{d}}(x,y) \ln p_\theta(x,y) \mathrm{d}x\mathrm{d}y + \iint p_{\mathrm{d}}(x,y) \ln p_{\mathrm{d}}(x,y) \mathrm{d}x\mathrm{d}y & \text{(S5)} \\ &= -\mathrm{E}[\ln p_\theta(x,y)] + C & \text{(S6)} \\ &\approx -\frac{1}{N} \sum_{i=1}^{N} \ln p_\theta(x_i, y_i^{\mathrm{true}}) + C & \text{(S7)} \end{aligned}$$

In the above formula, since p_{d} is known, the integral of the second term in Eq. S5 is a constant. Specifically, the true distribution of the data is

$$p_{\mathrm{d}}(x,y) = p_{\mathrm{d}}(y|x) p_{\mathrm{d}}(x) = \delta(y - y^{\mathrm{true}}(x)) p_{\mathrm{d}}(x). \tag{S8}$$

Therefore, the distribution of predicted values $p_\theta(y|x)$ can reasonably be assumed to be a Gaussian distribution:

$$p_\theta(x,y) = p_\theta(y|x) p_\theta(x) = \frac{1}{\sqrt{2\pi\sigma_\theta^2(x)}} \exp\left(-\frac{(y - \mu_\theta(x))^2}{2\sigma_\theta(x)} \right) \cdot p_\theta(x) \tag{S9}$$

Further,

$$\ln p_\theta(x,y) = -\frac{1}{2} \ln 2\pi\sigma_\theta^2(x) - \frac{(y - \mu_\theta(x))^2}{2\sigma_\theta(x)} + \ln p_\theta(x) \tag{S10}$$

is substituted into Eq. S7 and set σ_θ to be a constant, the KL divergence is restored to the original MSE loss function,

$$D_{\mathrm{KL}}(p_{\mathrm{d}} \| p_\theta) \sim \frac{1}{N} \sum_{i=1}^{N} (\mu_\theta(x_i) - y_i^{\mathrm{true}})^2. \tag{S11}$$

Now we introduce some data for which we only know the approximate range of affinity values, namely

$$\{(x_i, (-\infty, y_i^{\text{true}}))\}_{i=1}^{N}. \tag{S12}$$

Without any prior knowledge, y can be defaulted to be uniform:

$$p_{\text{d}}(y|x) = \begin{cases} \epsilon & \text{if } y < y^{\text{true}}(x), \\ 0 & \text{if } y \geq y^{\text{true}}(x). \end{cases} \tag{S13}$$

Therefore, the distribution of predicted values can be assumed to be a Sigmoid function,

$$p_\theta(x, y) = p_\theta(y|x)p_\theta(x) = \frac{1}{1 + \exp(-(y - \mu_\theta(x)))} \cdot p_\theta(x). \tag{S14}$$

And further,

$$\ln p_\theta(x, y) = -\ln(1 + \exp(-(y - \mu_\theta(x)))) + \ln p_\theta(x). \tag{S15}$$

Hence the KL divergence at this time is

$$D_{\text{KL}}(p_{\text{d}} \| p_\theta) \sim \frac{1}{N} \sum_{i=1}^{N} \ln(1 + \exp(\mu_\theta(x_i) - y_i^{\text{true}})), \tag{S16}$$

which is exactly the Softplus function. It is a contrastive loss function that will make the prediction results as low as possible below a certain given value, which is applicable in some sense to the data we generate through docking. The Softplus function is just one of many possibilities, and many other types of loss functions can also make the distribution of prediction scores close to Eq. S13, such as those shown in Table S22. To clearly compare the differences between these functions, Figure S16 plots their curves and the probability distributions of the predicted values caused by them.

From the plot of probability distributions, we can observe that the optimization goals of the two loss functions, Softplus and exp, are actually to infinitely distance themselves from specific anchor points. Even with the constraint of positive samples, the final prediction

Table S22: Several feasible contrastive loss functions

Name	Formula[a]
HalfMSE	$(y_i^{\text{pred}} - y_i^{\text{true}})^2 H(y_i^{\text{pred}} - y_i^{\text{true}})$
ReLU	$(y_i^{\text{pred}} - y_i^{\text{true}}) H(y_i^{\text{pred}} - y_i^{\text{true}})$
Softplus	$\ln(1 + \exp(y_i^{\text{pred}} - y_i^{\text{true}}))$
exp	$\exp(y_i^{\text{pred}} - y_i^{\text{true}})$

[a] $H(x)$ is the Heaviside step function.

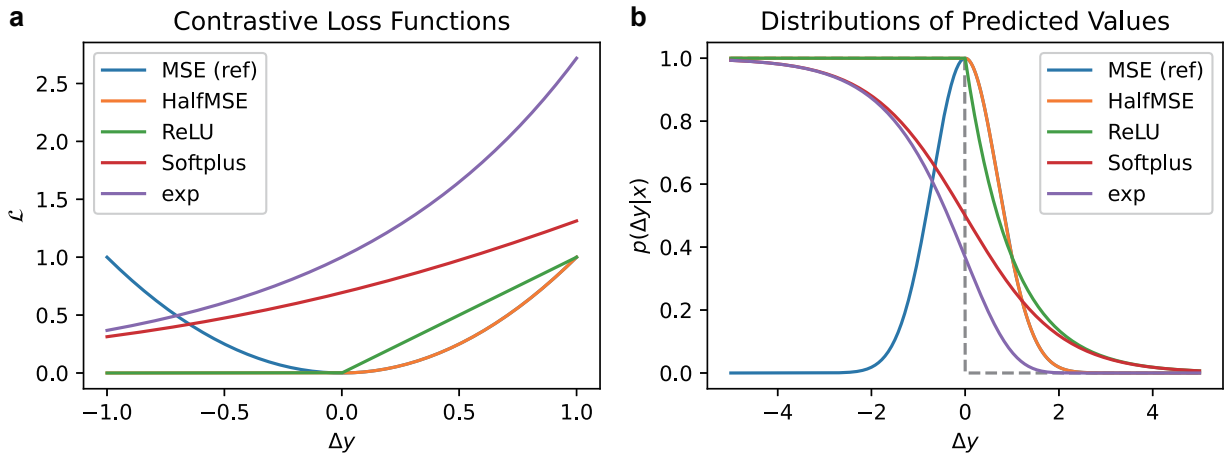

Figure S16: (**a**) Some contrastive loss functions that can make the predicted value of the optimized model less than a specified value. (**b**) The probability distribution of the predicted values these loss functions give rise to. Note that the MSE in the figure is for reference only.

of the model will tend to maintain a certain distance from the anchor points. As for the docking data we produced, we can only determine that their affinity scores are theoretically less than the corresponding scores of the crystal structure. As for how much less, we don't know. They could be very close or very far away. Therefore, Softplus and exp cannot fully meet our requirements.

On the other hand, HalfMSE and ReLU (a limit of Softplus) have a uniform maximum probability distribution on the left side of the anchor point, which means that the result of optimization is to make the predicted values fall in this area as much as possible. How far they need to be depends on the knowledge of the positive samples, which avoids artificially introducing a gap with the anchor point. They both meet our objectives, the only difference is a slight difference in the distribution on the right side of the anchor point. In this work, we chose HalfMSE as our contrastive loss function because it is close to the loss function used for the positive samples, provides faster optimization speed in the initial stage of training, and has a smoother transition when close to the anchor point.

3.4 Training

The hyperparameter settings used for training DeepRLI are shown in Table S23. The maximum number of training epochs is set to 1000 to provide a sufficiently large value that neither interferes with task execution nor prematurely terminates training under normal conditions, while ensuring timely termination in exceptional cases. Since an automatic learning rate decay strategy is employed, training automatically stops when the learning rate decreases below a critical threshold. In practice, training typically concludes between 300 and 450 epochs without ever reaching the 1000-epoch limit.

For training the DeepRLI model on a single "NVIDIA Tesla V100 SXM2 32GB" GPU, a batch size of 6 is used. This conservative batch size is chosen because each training data unit contains five fully connected complex graphs, and larger batch sizes would require significantly more GPU memory. The initial learning rate is set to 0.0002. If the validation

Table S23: Hyperparameters used in model training

Parameter Name	Parameter Value
maximum number of epochs	1000
batch size	6
initial learning rate	0.0002
reduction factor of learning rate	0.5
reduction patience of learning rate	15
minimum learning rate	10^{-6}
number of node features	39
number of edge features	39
hidden dimensions	64
number of attention heads	8

loss shows no improvement over 15 consecutive epochs, the learning rate is reduced by a factor of 0.5. Training terminates automatically once the learning rate falls below 10^{-6}.

Model-related hyperparameters are also listed in Table S23. The input dimensions for the first neural network layers of node and edge features depend on their respective feature dimensions, both set to 39. The hidden embedding dimensions for nodes and edges within the neural network are 64. The graph transformer module uses 8 attention heads.

Training remains stable with no observed overfitting, so the final epoch's model is directly adopted as the production model. The trained model parameters are publicly available on GitHub (see the "Code Availability" section in the main text). To instantiate the production model for inference, users can initialize the model with the provided hyperparameters and load the trained parameters via the released state dictionary file.

3.5 Computational overhead of DeepRLI

We assessed the computational overhead of DeepRLI using a server equipped with an "Intel Xeon Gold 6132 @ 2.60GHz" CPU and an "NVIDIA Tesla V100 SXM2 32GB" GPU. For the 0 Ligand Bias dataset, comprising 365 protein–ligand complexes, DeepRLI's average data preprocessing time is 0.441072 seconds per item, and the average inference time is 0.047680 seconds per item, resulting in a total average runtime of 0.488752 seconds per

complex. Under the same configuration, GenScore's average data preprocessing time is 1.725224 seconds per item, with an average inference time of 0.780646 seconds per item, leading to a total average runtime of 2.50587 seconds per complex. Consequently, DeepRLI operates five times faster than GenScore, demonstrating lower computational costs. When utilizing 100 parallel processes (feasible on a machine with dual AMD CPUs), DeepRLI can evaluate nearly 20 million distinct protein–ligand complexes in a day. In a virtual screening scenario involving the same protein with different ligands, the computational load for data preprocessing can be further reduced by pre-extracting the protein pockets, significantly enhancing processing speed.

References

(1) 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. Inf. Model.* **2019**, *59*, 895–913.

(2) Dunbar, J. B. J.; Smith, R. D.; Yang, C.-Y.; Ung, P. M.-U.; Lexa, K. W.; Khazanov, N. A.; Stuckey, J. A.; Wang, S.; Carlson, H. A. CSAR benchmark exercise of 2010: selection of the protein–ligand complexes. *J. Chem. Inf. Model.* **2011**, *51*, 2036–2046.

(3) Schindler, C. E. M. et al. Large-scale assessment of binding free energy calculations in active drug discovery projects. *J. Chem. Inf. Model.* **2020**, *60*, 5457–5474.

(4) Tran-Nguyen, V.-K.; Jacquemard, C.; Rognan, D. LIT-PCBA: an unbiased data set for machine learning and virtual screening. *J. Chem. Inf. Model.* **2020**, *60*, 4263–4273.

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

(6) 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. *J. Chem. Inf. Model.* **2014**, *54*, 1700–1716.

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

(8) Shen, C.; Zhang, X.; Hsieh, C.-Y.; Deng, Y.; Wang, D.; Xu, L.; Wu, J.; Li, D.; Kang, Y.; Hou, T.; Pan, P. A generalized protein–ligand scoring framework with balanced scoring, docking, ranking and screening powers. *Chem. Sci.* **2023**, *14*, 8129–8146.

(9) Chipot, C., Pohorille, A., Eds. *Free Energy Calculations: Theory and Applications in Chemistry and Biology*; Springer Series in Chemical Physics; Springer: Germany, 2007.

(10) Wang, L. et al. Accurate and reliable prediction of relative ligand binding potency in prospective drug discovery by way of a modern free-energy calculation protocol and force field. *J. Am. Chem. Soc.* **2015**, *137*, 2695–2703.

(11) 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.; Shaw, D. E.; Francis, P.; Shenkin, P. S. Glide: a new approach for rapid, accurate docking and scoring. 1. method and assessment of docking accuracy. *J. Med. Chem.* **2004**, *47*, 1739–1749.

(12) Halgren, T. A.; Murphy, R. B.; Friesner, R. A.; Beard, H. S.; Frye, L. L.; Pollard, W. T.; Banks, J. L. Glide: a new approach for rapid, accurate docking and scoring. 2. enrichment factors in database screening. *J. Med. Chem.* **2004**, *47*, 1750–1759.

(13) Durant, G.; Boyles, F.; Birchall, K.; Marsden, B.; Deane, C. M. Robustly interrogating machine learning-based scoring functions: what are they learning? *Bioinformatics* **2025**, *41*, btaf040.

(14) Yu, J.; Li, Z.; Chen, G.; Kong, X.; Hu, J.; Wang, D.; Cao, D.; Li, Y.; Huo, R.; Wang, G.; Liu, X.; Jiang, H.; Li, X.; Luo, X.; Zheng, M. Computing the relative binding affinity of ligands based on a pairwise binding comparison network. *Nat. Comput. Sci.* **2023**, *3*, 860–872.

(15) Salentin, S.; Schreiber, S.; Haupt, V. J.; Adasme, M. F.; Schroeder, M. PLIP: fully automated protein–ligand interaction profiler. *Nucleic Acids Res.* **2015**, *43*, W443–W447.

(16) Adasme, M. F.; Linnemann, K. L.; Bolz, S. N.; Kaiser, F.; Salentin, S.; Haupt, V. J.; Schroeder, M. PLIP 2021: expanding the scope of the protein–ligand interaction profiler to DNA and RNA. *Nucleic Acids Res.* **2021**, *49*, W530–W534.

(17) Clark, M.; Cramer III, R. D.; Van Opdenbosch, N. Validation of the general purpose tripos 5.2 force field. *J. Comput. Chem.* **1989**, *10*, 982–1012.

(18) Meng, E. C.; Lewis, R. A. Determination of molecular topology and atomic hybridization states from heavy atom coordinates. *J. Comput. Chem.* **1991**, *12*, 891–898.

(19) Chu, X.-J. et al. Discovery of [4-amino-2-(1-methanesulfonylpiperidin-4-ylamino)pyrimidin-5-yl](2,3-difluoro-6-methoxyphenyl)methanone (R547), a potent and selective cyclin-dependent kinase inhibitor with significant in vivo antitumor activity. *J. Med. Chem.* **2006**, *49*, 6549–6560.

(20) Morgan, D. O. Cyclin-dependent kinases: Engines, clocks, and microprocessors. *Annu. Rev. Cell. Dev. Biol.* **1997**, *13*, 261–291.

(21) Malumbres, M.; Barbacid, M. To cycle or not to cycle: A critical decision in cancer. *Nat. Rev. Cancer* **2001**, *1*, 222–231.

(22) Pantoom, S.; Vetter, I. R.; Prinz, H.; Suginta, W. Potent family-18 chitinase inhibitors:

x-ray structures, affinities, and binding mechanisms. *J. Biol. Chem.* **2011**, *286*, 24312–24323.

(23) Nichols, D. A.; Jaishankar, P.; Larson, W.; Smith, E.; Liu, G.; Beyrouthy, R.; Bonnet, R.; Renslo, A. R.; Chen, Y. Structure-based design of potent and ligand-efficient inhibitors of CTX-M class A -lactamase. *J. Med. Chem.* **2012**, *55*, 2163–2172.

(24) McKeown, M. R.; Shaw, D. L.; Fu, H.; Liu, S.; Xu, X.; Marineau, J. J.; Huang, Y.; Zhang, X.; Buckley, D. L.; Kadam, A.; Zhang, Z.; Blacklow, S. C.; Qi, J.; Zhang, W.; Bradner, J. E. Biased multicomponent reactions to develop novel bromodomain inhibitors. *J. Med. Chem.* **2014**, *57*, 9019–9027.

(25) Belkina, A. C.; Denis, G. V. BET domain co-regulators in obesity, inflammation and cancer. *Nat. Rev. Cancer* **2012**, *12*, 465–477.

(26) Lin, H.; Wang, S. Datasets for training and inference of DeepRLI. 2024; DOI: 10.5281/zenodo.11116386.