Supplementary Information for

DiffBindFR: An SE(3) Equivariant Network for Flexible Protein-Ligand Docking

Jintao Zhu[1, †], Zhonghui Gu[2, †], Jianfeng Pei[1, *], Luhua Lai[1,2,3, *]

[1] Center for Quantitative Biology, Academy for Advanced Interdisciplinary Studies, Peking University, Beijing, 100871, China
[2] Peking-Tsinghua Center for Life Sciences, Academy for Advanced Interdisciplinary Studies, Peking University, Beijing, 100871, China
[3] BNLMS, College of Chemistry and Molecular Engineering, Peking University, Beijing, 100871, China

†Equal Contribution.

*Corresponding authors: lhlai@pku.edu.cn or jfpei@pku.edu.cn

This file contains the following information:

1. The protocol of building CD cross-dock test set

For the complexes in the CD test set, we employed the following methodology to search for their Apo states and other Holo states:

1. **ApoRef Subset:** We directly utilize the dataset processed by Zhang et al[1].
2. **CASF2016 Subset:** This subset contains 57 different proteins, each with five ligand-bound Holo states. We initially search for the Apo states of these 57 target proteins in the ApoBind database[2]. In cases where the Apo state is not available, we employed the AHoJ tool[3] to search for it, and ensure that each Holo state has at most three Apo states. Then we pair the Apo structures and other Holo structures in each target with the Holo structure for the cross-docking experiment.
3. **Ensemble CDK2, EGFR, and FXA Subsets:** These subsets consist of complex structures of three target proteins: CDK2, EGFR, and FXA. We search the PDB database for Holo structures using the Uniprot ID of these proteins and analyze whether they contain drug-like ligands in the orthosteric site to differentiate between Apo and Holo states. For CDK2, Apo PDB id: 1FIN, Uniprot ID: P24941; For EGFR, Apo PDB id: 7A2A, Uniprot ID: P00533; For FXA, Apo PDB id: 1EZQ, Uniprot ID: P00742.
4. **DUDE27-HoloEns Subset**: In this subset, we used the dataset (herein named DUDE27-AF2) reported by Zhang et al.[4] as the reference set of the Holo-Holo cross-dock test set. This test set contains Holo, AF2 modeled structures (referred as truncated AF2 structures in the original paper[4]) and IFD-MD[5] refined AF2 modeled structures of 27 targets from DUD-E[6] (see details in the Supplemental Table. S10). We firstly queried the 27 Holo structures to Proteins*Plus* web server[7], and used SIENA[8] module, a fully-automated protein binding site ensemble search tool, to search homologous Holo binding sites with the following general settings: (1) binding site radius is set to 5 Å, (2) Holo structures only, (3) the minimum site identity is set to 1.0, (4) complete residues

only, (5) no mutations in the global alignment, (6) remove sites with ligand duplicates, (7) the size of the interaction-based ensemble reduction is 10, (8) No PDB id found in our training set, (8) other parameters follows Screening Mode. Further, additional parameters (such as resolution threshold is set to 3 Å and allow mutations in the global alignment but not in binding site) were optimized to ensure that at least two Holo structures can be searched for each target as much as possible. Note that the query Holo structures were excluded from the searched Holo set. To control the quadratic growth of pairs ($O(n^2)$) while preserving interaction diversity, we limited the Holo-Holo cross-docking for each target to at most five Holo structures with the slightest backbone RMSD (e.g., seven targets have more than five searched Holo structures). See details in the supplemental Table. S2 about the DUDE27-HoloEns subset.

5. **GPCR-AF2 Subset:** Compiled by Karelina et al.[9], this subset comprises 66 GPCR extracellular domain-ligand complex structures published in the PDB after April 30, 2018. It includes 18 different GPCR proteins (17 class A, 1 class B). We predict the structures of these 18 GPCRs using AlphaFold2[10], restricting the structure templates to those dated before April 30, 2018. The structure with the highest pLDDT score is selected and further optimized using amber relax[11].

Our methodology for processing the structures obtained from our search is as follows: Taking ligand *i* as an example, its experimental Holo state structure is denoted as Holo *i*. We utilize the align_binding_sites module of the Schrödinger software suite[12] to overlay the Holo *i* structure with either the Apo state structure or other Holo state structures bound to different ligands (non-Holo *i*), using the default parameters -cutoff 5 -dist 5. The resultant overlaid structure (non-Holo *i*) serves as the target protein structure for cross-docking input of ligand *i*.

For the ligand small molecules, we prepare them using the prepwizard module of the Schrödinger software, which involves converting the PDB format to SDF format. In the post-processing phase, we meticulously examine the cross-docking structures for potential clashes. If a severe clash exists between the ligand and the protein, we exclude that particular sample from our analysis.

2. System preparation

To avoid introducing bias toward the crystal ligand, proteins and ligands were first separated from the PDB complex structures, then prepared independently using the prepwizard module of the Schrödinger software with default settings. The protein preparation pipeline included removing water molecules, adding hydrogen atoms, filling missing side chain atoms, assigning bond orders, and optimizing the H-Bond network. A restrained minimization was performed with the fixed backbone, optimizing hydrogen atoms using the OPLS_2005 force field[13] to preserve the conformations of the binding sites as much as possible. PROPKA[14, 15] and Epik were used to assign the protonation and ionization states of the proteins, respectively, at pH=7.0. Ligands were treated similarly according to the above protein preparation scheme, respecting chiralities from the input geometry based on the crystal structure. Meanwhile, we carefully checked the prepared ligand structures to ensure they could be readable for RDKit[16] as much as possible. The prepared proteins and ligands were further processed according to the official documentation of each docking program for suitable docking inputs.

3. Baseline methods

3.1 Vina

AutoDock Vina[17] is a widely-used traditional docking method. Ligands were converted from SDF format to PDBQT format by the mk_prepare_ligand.py script from Meeko v0.5[18]. Protein PDBQT files were generated by the prepare_receptor script with the additional argument -A ‘checkhydrogens’ in ADFR Suite 1.0. We defined the box using the center of the ligand present in the crystal structure, setting the box dimensions to 24×24×24 $Å^3$. The ‘exhaustiveness’ parameter in Vina was set to 32, producing up to 10 poses for each docking run. Docking was repeated running 40 times with different random seeds to get the top-ranked pose.

3.2 Smina

Smina[19] improves AutoDock Vina with a new scoring function and is more easy-to-use. The PDBQT file preparation, box construction and the sampling strategy were the same from the aforemetioned baseline method AutoDock Vina.

3.3 LinF9

LinF9[20] improves Autodock Vina with a new scoring function and is more user-friendly. The PDBQT file preparation, box construction and the sampling strategy were the same from the aforemetioned baseline method AutoDock Vina.

3.4 Gold

Gold[21] is another widely-used traditional docking method. The binding sites were defined as pocket residues within radius 12.5 Å around the crystal ligand. The settings used were rescore function ‘plp’, autoscale 10, and early termination off. The docking performance was taken from Buttenschoen et al. reported[22].

3.5 VinaFlex

AutoDock Vina also supports flexible docking with movable side chains[23]. However, it requires the explicit designation of the side chains allowed to move and can support up to 14 flexible residues. Before each docking attempt, we randomly selected up to 14 residues within the defined 24×24×24 Å^3 box to act as the flexible residues. The ligand preparation was consistent to AutoDock Vina, but protein preparation required an additional scheme for preparing flexible residues. Here, we used a python script prepare_flexreceptor.py available at https://github.com/ccsb-scripps/AutoDock-Vina/tree/develop/example/autodock_scripts to obtain two PDBQT files, one for rigid part and the other for flexible side chains. The 'exhaustiveness' parameter was set to 16. Each docking run generated up to 10 poses, and this docking process was repeated running 40 times using different random seeds to get the top-ranked pose.

3.6 rDock

rDock[24] is another traditional docking method. The protein input files for rDock are in Mol2 format which can be converted from Schrödinger Mae format files by structconvert module of Schrödinger software. The ligand input files are in SDF format, directly taken from prepared ones in the Section 2 in the supplementary information. The box construction and the sampling strategy were the same from the aforemetioned baseline method AutoDock Vina. Otherwise, functional groups, specifically -OH and -NH3+, located within 3 Å of the ligand on the pocket residues were allowed to move. Docking was repeated running 40 times with different random seeds to get the top-ranked pose.

3.7 Glide

Glide[25] is a powerful commercial docking method. The rigid receptor docking was executed using the Glide-SP docking method in the Schrodinger software suite. The protein and ligand preparation protocol has been described in the Section 2 in the supplementary information. For the generation of grid files, the parameter

‘INNERBOX’ was set to 15 and ‘UTERBOX’ was set to 30, with all other parameters as default. Each docking run produced a maximum of 10 poses, and the docking was repeated running 40 times to get the top-ranked pose.

3.8 TankBind

TankBind[26] is a recently developed deep learning-based method. The protocol and setting followed the official tutorial, which is available at https://github.com/luwei0917/TankBind/blob/main/examples/testset_evaluation_cleaned.ipynb. Instead of using the P2Rank prediction for pocket localization, the model utilizes the center of the ligand from the crystal structure, with all other parameters set to their default values. Since this method reconstructs ligand coordinates from the predicted distance matrix of complex, it can only generate a single pose for the ligand.

3.9 EDM-Dock

EDM-Dock[27] is a deep learning-based method sharing similar algorithm with TankBind. The protocol in the README file in the EDM-Dock repository (https://github.com/MatthewMasters/EDM-Dock) were used for docking. The box was defined as a 22.5×22.5×22.5 Å^3 cube. Extra energy minimization was performed for the single ligand pose predicted by EDM-Dock.

3.10 KarmaDock

KarmaDock[28] is a recently developed deep learning-based regression model which predicts ligand coordinates directly in the Euclidean space. Following the protocol from the KarmaDock article[28], we reproduced its reported results on the CASF2016 test set (Supplemental Table. S3), showing that we successfully re-trained the original KarmaDock. For fair comparison with our model, we further re-trained KarmaDock using the PDBbind time-split training set without any artificial intervention. KarmaDock docking was run with its default parameters.

Additionally, we augmented the KarmaDock model with a ResNet module to predict the side chain torsion angles of the binding pocket, resulting in a refined model named KarmaDock-sc (see Supplemental Fig. S4).

3.11 DiffDock

DiffDock[29] is a blind-docking method based on diffusion generative model. Although it's not fair to compare DiffDock with pocket-docking methods, we still evaluate its performance to reflect the defect of ignoring physical plausibility of these deep learning-based methods. Each generation of ligand poses was repeated running 40 times, and the generated poses were ranked by DiffDock confidence model. Again, the docking performance was taken from Buttenschoen et al. reported[22].

4. The composition of CD cross-dock test set

Table S1. Numbers of cross-dock pairs in CD test set. [α]

Subset	Type	No. pfam	No. Apo	No. Holo	No. Crossdock
Ensemble-CDK2	Apo-Holo	1	34	339	11317
Ensemble-EGFR	Apo-Holo	1	1	72	67
Ensemble-FXA	Apo-Holo	1	4	109	436
ApoRef	Apo-Holo	32	64	293	548
CASF2016	Apo-Holo;Holo-Holo	57	338	285	1760
DUDE27-HoloEns	Holo-Holo	27	0	93	268
GPCR-AF2	AF2 Structure-Holo	1	18	66	66

[α] No. pfam denotes the number of pfam for target proteins in each subset. No. Apo and No. Holo denotes the number of protein Apo states (without drug-like ligand binding) and protein Holo states in each subset. No. Crossdock denotes the total number of Apo-Holo pairs and Holo-Holo pairs in each subset. For GPCR-AF2 subset, the AlphaFold2 predicted GPCR structures are counted as Apo proteins. DUDE27-HoloEns is a subset that only comprises of Holo-Holo cross-dock pairs, and the related details can be found in Table. S2.

Table S2. Details about the searched Holo structures in DUDE27-HoloEns subset.

Target	PDB code	PDB chains	Active site identity	Backbone RMSD[α]	All atom RMSD[α]
dpp4	2AJ8	A	1.00	0.24	0.49
dpp4	5LLS	A	1.00	0.25	0.34
dpp4	2BUC	A	1.00	0.26	0.62
ptn1	8SKL	A	1.00	0.22	0.61
ptn1	7MM1	A	1.00	0.25	0.66
ptn1	7FQU	A	1.00	0.28	0.86
aces	7AIS	A	1.00	0.17	0.41
aces	4TVK	A	1.00	0.27	0.48
aces	6H12	A	1.00	0.31	0.54
aces	5EHX	A	1.00	0.33	0.76
aces	1GQR	A	1.00	0.33	0.79
braf	5ITA	A	1.00	1.29	1.38
braf	7M0X	A	1.00	1.93	2.13
braf	7P3V	A	1.00	1.95	2.00
braf	6P3D	A	1.00	2.38	2.85
braf	6N0Q	A	1.00	2.54	2.81

vgfr2	6GQO	A	1.00	2.01	1.93
vgfr2	6XVK	A	1.00	2.01	1.91
akt2	3E87	A	1.00	0.50	1.38
akt2	1O6K	A	1.00	0.52	1.45
akt2	2UW9	A	1.00	0.58	1.57
akt2	2JDR	A	1.00	0.69	1.06
tgfr1	5FRI	A	1.00	0.42	0.86
tgfr1	2WOT	A	1.00	0.44	0.82
tgfr1	4X0M	A	1.00	0.70	0.93
mapk2	1NY3	A	1.00	0.64	0.97
mapk2	6T8X	A	1.00	0.77	0.92
mapk2	3KA0	A	1.00	4.31	3.21
tryb1	4MPU	A	1.00	0.15	0.80
tryb1	4MPW	A	1.00	0.16	0.70
tryb1	4MPV	A	1.00	0.17	0.58
tryb1	5F03	A	1.00	0.18	0.99
try1	2AYW	A	1.00	0.11	0.30
try1	3A7W	A	1.00	0.28	0.74
thrb	6YSX	H	1.00	0.24	0.51
thrb	2ZG0	H	1.00	0.26	0.45
thrb	3U9A	H	1.00	0.35	0.50
thrb	6ZUW	H	1.00	0.39	0.81
thrb	6ZV8	H	1.00	0.42	0.72
ppard	7VWG	A	1.00	0.26	0.63
ppard	5U43	A	1.00	0.37	0.84
ppard	1GWX	A	1.00	0.41	0.92
ppard	5U46	A	1.00	0.45	0.79
ppard	7WGN	A	1.00	0.52	0.92
pparg	7WGO	A	1.00	0.39	0.77
pparg	6MS7	A	1.00	0.40	1.43
pparg	2VST	A	1.00	0.41	1.20
pparg	2HWR	A	1.00	0.42	0.92
pparg	6ZLY	A	1.00	0.43	0.73
fa10	3KQB	A	1.00	0.26	0.65
fa10	3M37	A	1.00	0.32	0.79
fa10	4Y71	A	1.00	0.40	0.97
cdk2	3SW7	A	1.00	0.44	1.33
cdk2	3QRT	A	1.00	0.54	1.24
mk10	2ZDU	A	1.00	0.26	1.12
mk10	1PMV	A	1.00	0.58	1.02
mk10	2O0U	A	1.00	0.66	1.49
rxra	6STI	A	1.00	0.21	0.63
rxra	7UW2	A	1.00	0.24	0.76
rxra	7B9O	A	1.00	0.24	0.71

rxra	2P1T	A	1.00	0.25	0.66
rxra	7NKE	A	1.00	0.25	1.01
mk14	3ZSH	A	1.00	1.78	1.61
mk14	5N65	A	1.00	1.99	1.96
gria2	3KGC	B	1.00	0.16	0.24
gria2	1GR2	A	1.00	1.38	1.50
gria2	2AIX	A	1.00	1.74	1.78
egfr	8A27	A	1.00	0.51	1.01
egfr	7KXZ	A	1.00	1.20	1.95
egfr	8F1Z	A	1.00	1.21	1.93
egfr	7U99	A	1.00	1.28	1.94
egfr	8DSW	A	1.00	1.58	2.28
igf1r	1K3A	A	0.86	0.74	0.97
igf1r	1JQH	A	0.95	3.83	4.23
ampc	6WHF	B	1.00	0.24	0.71
met	2RFS	A	1.00	2.05	2.71
met	2WKM	A	1.00	2.10	2.72
met	7B3Q	A	1.00	2.17	2.81
bace1	4B0Q	A	1.00	0.18	0.70
bace1	6UVP	A	1.00	0.20	0.37
bace1	4FSL	A	1.00	0.33	0.53
bace1	6JT4	A	1.00	1.16	1.24
hs90a	3WQ9	A	1.00	0.49	0.58
hs90a	5VYY	A	1.00	0.86	1.09
hs90a	7UR3	A	1.00	1.38	1.76
hs90a	3T0Z	A	1.00	1.46	1.46
fabp4	7FVY	A	1.00	0.40	0.69
fabp4	7FVV	A	1.00	0.45	0.76
fabp4	7FWZ	A	1.00	0.45	0.72
fabp4	7FZJ	A	1.00	0.52	0.75
fabp4	7FYT	A	1.00	0.54	0.95
ital	4IXD	A	1.00	0.41	1.06
ital	3BQM	B	1.00	0.58	1.32

[a] The RMSD is calculated between the query Holo structure from DUD-E and searched Holo structure from the SIENA tool.

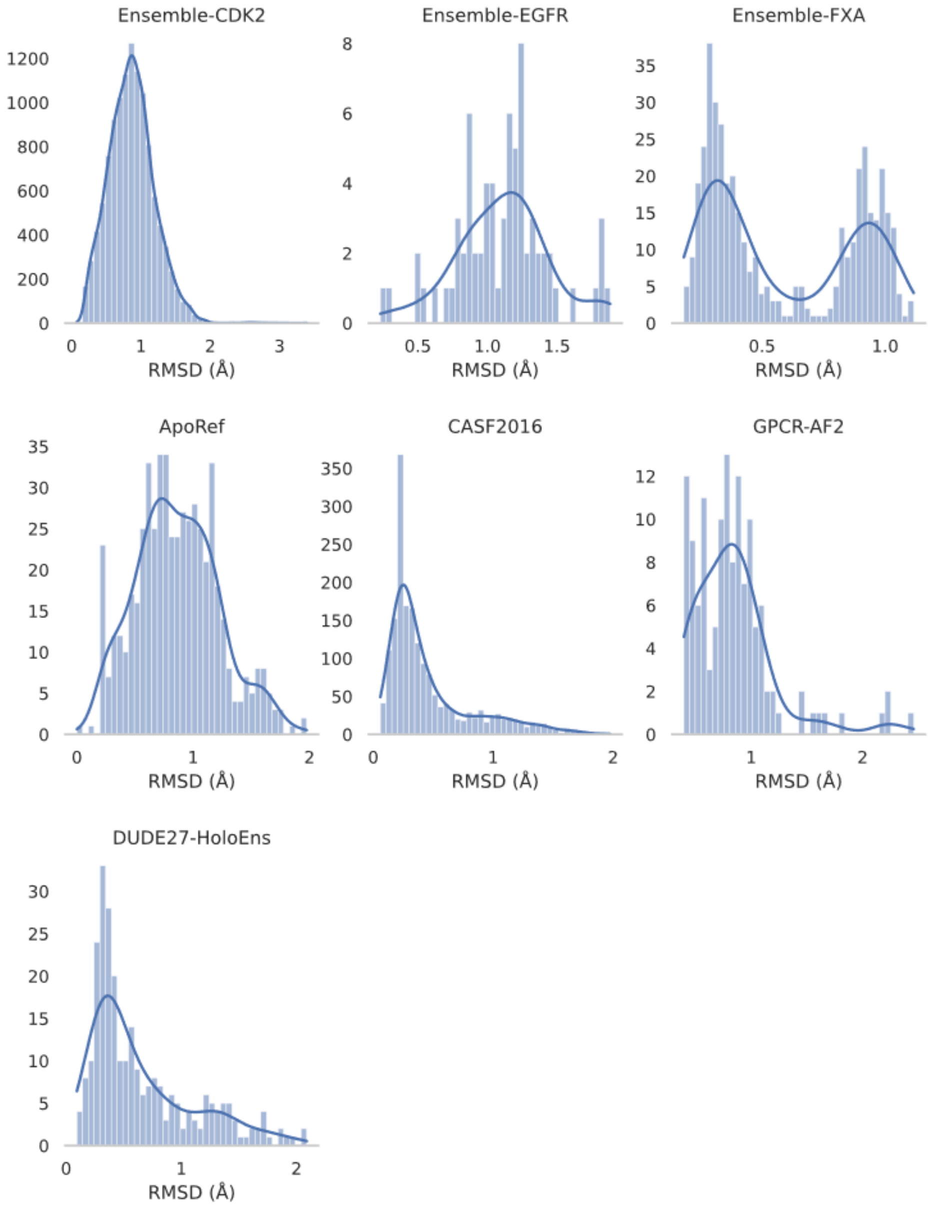

Fig. S1. Pocket backbone Cα RMSD distribution of cross-dock pairs on each subset from CD test set.

5. The details of MDN confidence model

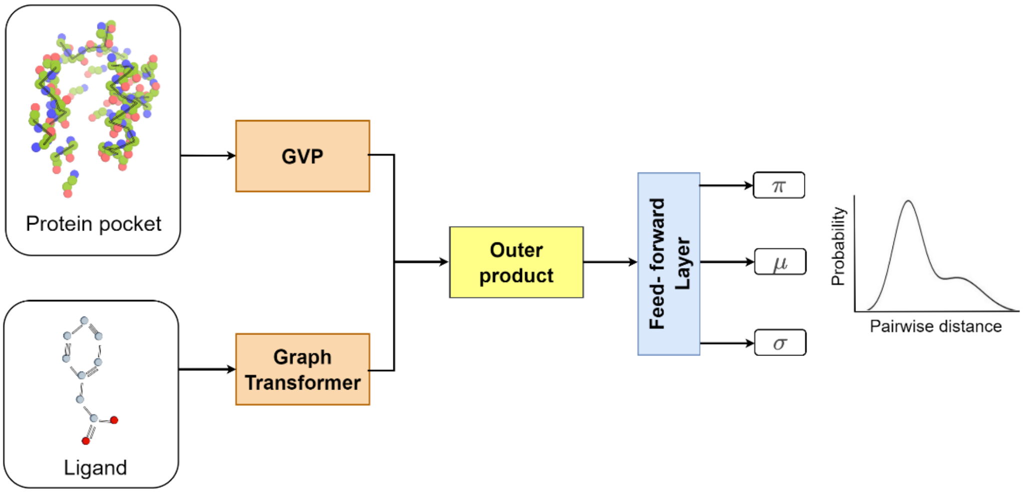

Fig. S2. Architecture of the MDN (mixture density network) confidence model.

The protein backbone's structural encoding is accomplished via the GVP (Geometric Vector Perceptron) module[30], resulting in node embeddings for each pocket residue, denoted as $\mathbf{v}^{\mathrm{p}}$. Concurrently, the ligand graph undergoes encoding through a Graph Transformer module[31], yielding node embeddings for each ligand atom, symbolized as $\mathbf{v}^{\mathrm{l}}$. This process is followed by the computation of ligand-residue pairwise distance embeddings, achieved via the outer product method. Subsequently, a feed-forward neural layer is employed to predict the parameters of a Gaussian mixture model[32], which characterizes the distribution of each pairwise distance.

The model is trained to minimize the loss function, as depicted in equation (1). This function encompasses multiple components: $\mathcal{L}_{\mathrm{MDN}}$ represents the loss associated with the mixture density network; $\mathcal{L}_{\mathrm{atoms}}$ and $\mathcal{L}_{\mathrm{bonds}}$ denote the cross-entropy cost functions for predicting atom and bond types, respectively, which serve as auxiliary tasks. Notably, $\mathcal{L}_{\mathrm{MDN}}$ is designed to minimize the negative log-likelihood of $d_{r,s}$, which signifies the minimum distance between the atoms of residue r and ligand atom s. This distance is calculated using a mixture model comprising $K=10$ Gaussian distributions, parametrized by $\pi_{r,s}$, $\mu_{r,s}$ and $\sigma_{r,s}$, as predicted by the model (refer to Equation (2)). The final confidence score is computed using the equation (3).

$$\mathcal{L}_{\mathrm{total}} = \mathcal{L}_{\mathrm{MDN}} + \mathcal{L}_{\mathrm{atoms}} + \mathcal{L}_{\mathrm{bonds}} \qquad (1)$$

$$\mathcal{L}_{\mathrm{MDN}} = -\log P\left(d_{r,s}\left|\mathbf{v}_{\mathbf{r}}^{\mathrm{p}}, \mathbf{v}_{\mathbf{s}}^{\mathrm{l}}\right.\right) = -\log \sum_{k=1}^{K} \pi_{r,s,k}\mathcal{N}\left(d_{r,s}\left|\mu_{\mathrm{r,s,k}}, \sigma_{r,s,k}\right.\right) \tag{2}$$

$$U_{(x)} = -\sum_{r=1}^{R}\sum_{s=1}^{S}\log P\left(d_{r,s}\left|\mathbf{v}_{\mathbf{r}}^{\mathrm{p}}, \mathbf{v}_{\mathbf{s}}^{\mathrm{l}}\right.\right) = -\mathrm{Score} \tag{3}$$

6. The details about re-training KarmaDock

Table S3. The splitting methods on PDBbind V2020 dataset. α

	Time-split	MLSF-split
Splitting metrics	Uploading time & ligand overlap	Sequence similarity
Training set size	16379	17242
Validation set size	968	1916
Test set size	363	285
Protein sequence similarity	0.484	1.00

αThe time-split method follows the work of EquiBind, where 363 complex structures from PDBbind 2020 dataset uploaded later than 2019 serve as test set. After removing ligands that exist in the test set, the remaining 16739 structures are used for training and 968 structures are used for validation. The MLSF-split method is used by the work of KarmaDock, where 90% of PDBbind general set serve as training set, 10% of PDBbind general set serve as validation set, and CASF2016[33] serves as test set. **Protein sequence similarity** represents protein sequence similarity between test set and training & validation set. As is analyzed by Zhang et al.[28], MLSF-split method causes all the protein sequences in test set existing in training set, while time-split method results in a more reasonable protein similarity between test set and training & validation set.

Table S4. The performance of KarmaDock without conformation correction from various scenarios. β

Result from	Dataset		success rate
	training set	test set	
Published data	MLSF-split training set	MLSF-split test set	89.1%
Released model	MLSF-split training set	MLSF-split test set	81.2%
Re-trained model	MLSF-split training set	MLSF-split test set	87.4%
Published data	Time-split training set	Time-split test set	56.2%
Released model	MLSF-split training set	Time-split test set	55.7%
Re-trained model	Time-split training set	Time-split test set	42.7%

βWe have re-trained KarmaDock in both methods for PDBbind dataset splitting. **Publish** represents the published success rate in the corresponding test set from KarmaDock article[28]; **Release** represents the success rate of released KarmaDock model by Zhang et al. from their github repository (https://github.com/schrojunzhang/KarmaDock); **Re-train** represents the success rate of KarmaDock model re-trained by us. On MLSF-split dataset, we find that we can reproduce the performance of KarmaDock published in the article (89.1% vs 87.4%). On time-split dataset, using the same training protocol, we cannot reproduce the published docking success rate of KarmaDock on time-split test set (56.2% vs 42.7%), but we find that the released model trained on MLSF-split training set have similar docking success rate with the article published data (56.2% vs 55.7%).

7. Supplementary information in PDBbind time-split test set

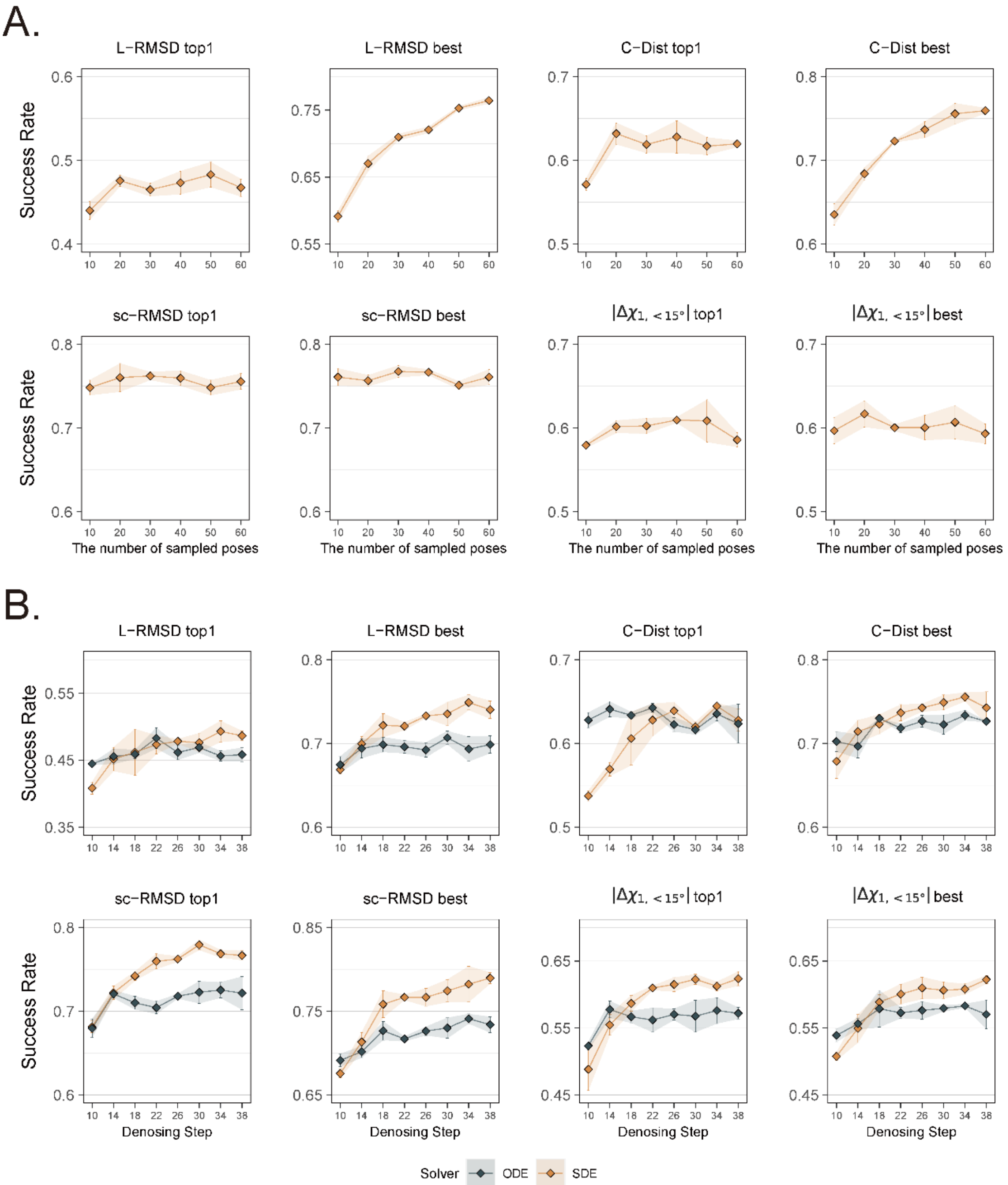

Fig. S3. Ablation experiments of DiffBindFR network sampling steps (A) and denoising paradigm (B, ODE/SDE) and denoising steps (B) on PDBbind time-split test set. The top-1 ligand poses used for evaluation are selected by MDN confidence model, and have not undergone any local optimization.

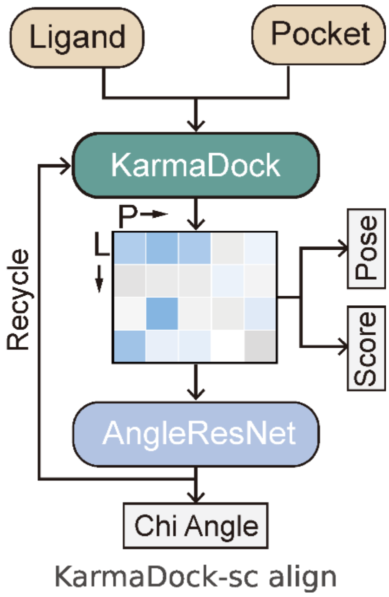

Fig. S4. Architecture of side chain torsion angle prediction module in KarmaDock-sc. The module is implemented based on ResNet, and the cosine and sine of a torsion angle is predicted following AlphaFold2[10].

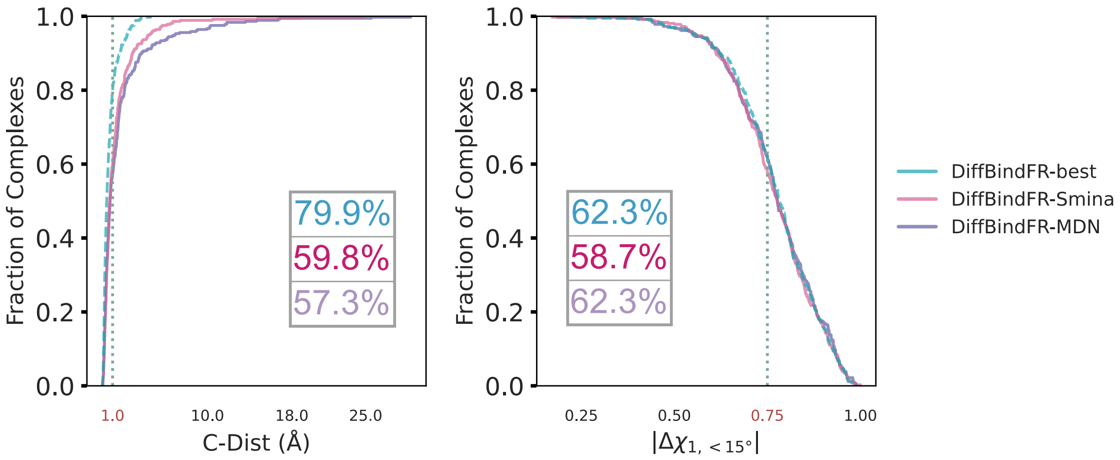

Fig. S5. Distributions of DiffBindFR on PDBbind time-split test set for C-Dist and $|\Delta\chi_{1,\ <15°}|$. C-Dist denotes ligand centroid distance, and $|\Delta\chi_{1,\ <15°}|$ denotes proportion of pocket residues with $|\Delta\chi_1| < 15^{\circ}$.

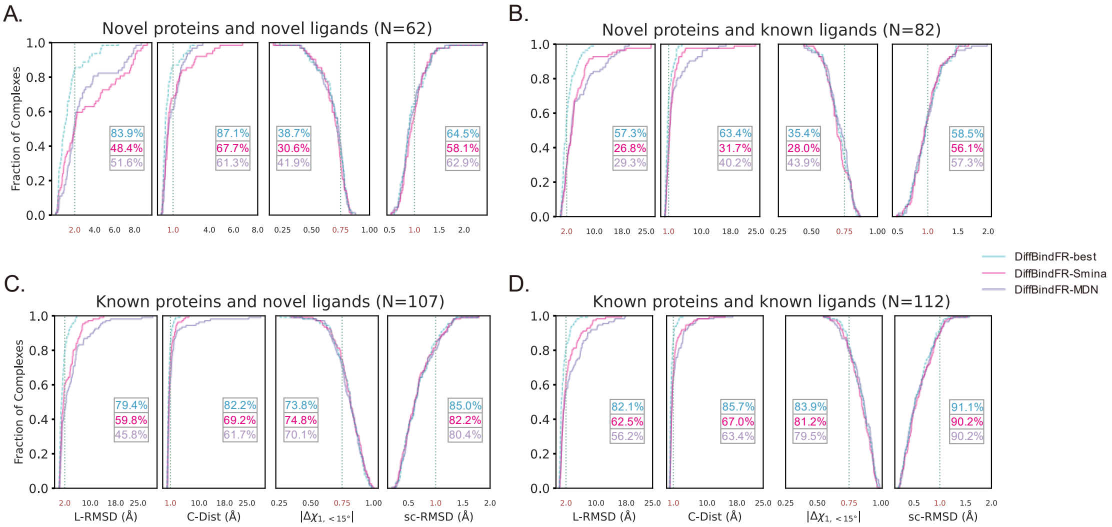

Fig. S6. Evaluation of DiffBindFR generalizability on PDBbind time-split test set. A protein is considered novel based on no Uniprot ID overlap, and a ligand based on a 0.5 Tanimoto similarity coefficient cut off relative to PDBbind time-split training set (using 1024 bit RDKit fingerprints[16]).

8. Supplementary information in Posebusters test set

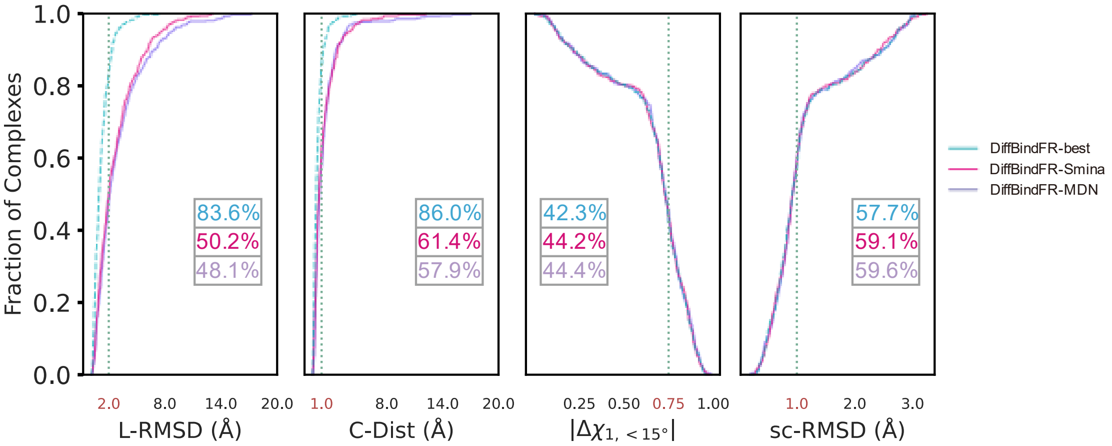

Fig. S7. Performance of DiffBindFR on Posebusters test set. For each complex, 40 poses are generated

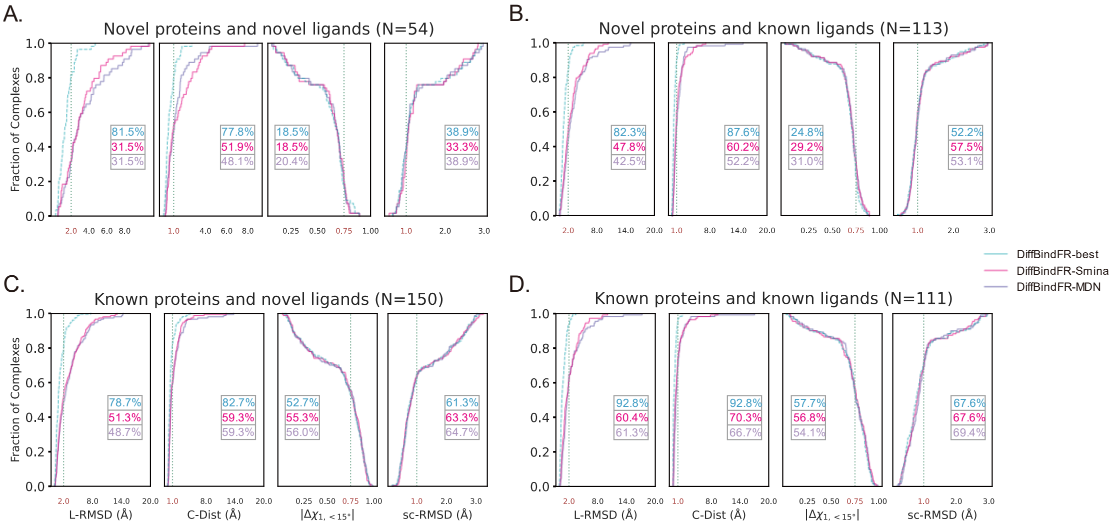

Fig. S8. Evaluation of DiffBindFR generalizability on Posebusters test set. A protein is considered novel based on a 40% sequence identity cut off, and a ligand based on a 0.5 Tanimoto similarity coefficient cut off relative to PDBbind time-split training set (using 1024 bit RDKit fingerprints[16]).

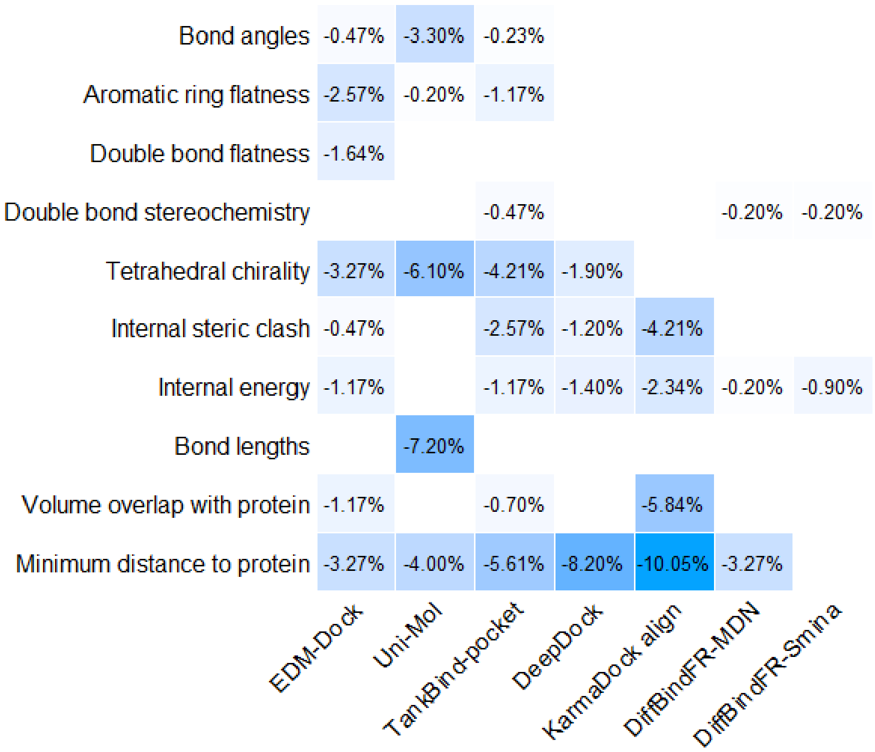

Fig. S9. The invalid rate of the top-1 poses with L-RMSD < 2 Å generated by various deep learning-based methods on different Posebusters terms. The invalidity of each term is evaluated by Posebusters suite, and the number in each box from the plot represents the proportion of poses fails in each term.

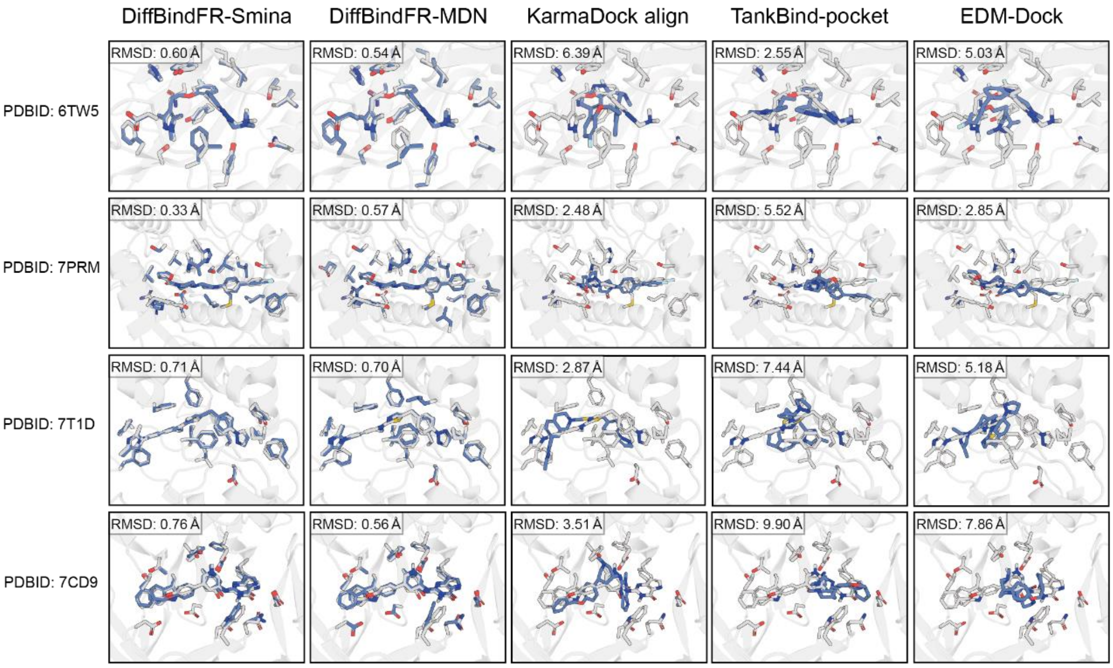

Fig. S10. The binding poses generated by various deep learning-based methods from Posebusters test set. Four cases (PDB id: 6TW5, 7PRM, 7T1D, 7CD9) are visualized, where conformations in gray and blue represent crystal structures and docking structures, respectively.

Table S5. Posebusters terms that various deep learning-based methods fail in. [α]

	6TW5	7PRM	7T1D	7CD9
EDM-Dock	aromatic_ring_flatness, internal_steric_clash, internal_energy, volume_overlap_with_protein, minimum_distance_to_protein	bond_angles, aromatic_ring_flatness, minimum_distance_to_protein	aromatic_ring_flatness, internal_steric_clash, volume_overlap_with_protein, minimum_distance_to_protein	aromatic_ring_flatness, tetrahedral_chirality, volume_overlap_with_protein, minimum_distance_to_protein
KarmaDock Align	internal_steric_clash, internal_energy, volume_overlap_with_protein, minimum_distance_to_protein	volume_overlap_with_protein, minimum_distance_to_protein	volume_overlap_with_protein, minimum_distance_to_protein	volume_overlap_with_protein, minimum_distance_to_protein
TankBind-pocket	internal_steric_clash, internal_energy, minimum_distance_to_protein	internal_steric_clash, internal_energy, volume_overlap_with_protein, minimum_distance_to_protein	aromatic_ring_flatness, internal_steric_clash, internal_energy, volume_overlap_with_protein, minimum_distance_to_protein	tetrahedral_chirality, internal_steric_clash, internal_energy, bond_lengths, volume_overlap_with_protein, minimum_distance_to_protein
DiffBindFR-mina	-	-	-	-
DiffBindFR-MDN	-	-	-	-

[α]Four cases from Fig. S9 are analyzed through Posebusters suite[22] to check their pose validity.

9. Supplementary information in CD test set

Table S6. L-RMSD of various methods on the subsets of CD test set. $^{\alpha}$

Method		**Ensemble-CDK2**		**Ensemble-EGFR**		**Ensemble-FXA**	
		RMSD Mean	**RMSD Median**	**RMSD Mean**	**RMSD Median**	**RMSD Mean**	**RMSD Median**
Traditional rigid receptor docking methods	Vina	6.12±2.63	6.26	6.37±2.34	6.90	4.35±3.17	3.02
	LinF9	6.12±2.50	6.31	6.31±2.43	6.63	3.88±2.90	2.96
	Smina	6.15±2.65	6.28	6.33±2.53	6.90	4.17±3.16	2.70
	Gnina	5.67±2.63	5.78	6.02±2.69	6.29	3.83±2.99	2.44
	Glide	4.81±2.60	5.21	6.87±3.63	7.67	3.49±2.72	2.31
Traditional flexible docking methods	VinaFlex	8.04±2.92	7.88	9.03±2.47	9.45	9.19±2.22	9.42
	rDock	4.61±2.76	4.62	5.45±2.98	4.70	3.29±2.66	2.28
Deep learning-based docking methods	TankBind-pocket	2.17±1.85	1.62	2.21±0.96	1.82	1.73±0.96	1.50
	EDM-Dock	2.62±1.27	2.32	2.95±1.05	2.66	2.68±0.98	**0.45**
	KarmaDock Align	1.89±1.20	1.58	2.65±2.06	1.93	2.43±0.92	2.30
	DiffBindFR-Smina	2.31±1.80	1.73	3.44±2.53	2.57	1.54±1.14	1.30
	DiffBindFR-MDN	**1.85**±1.34	**1.48**	**2.58**±2.15	**1.80**	**1.42**±0.70	1.35

Table S6. (continued)

Method		**ApoRef**		**CASF2016**		**GPCR-AF2**	
		RMSD Mean	**RMSD Median**	**RMSD Mean**	**RMSD Median**	**RMSD Mean**	**RMSD Median**
Traditional rigid receptor docking methods	Vina	6.98±3.35	7.22	4.93±3.62	4.46	7.42±3.99	7.77
	LinF9	6.59±3.08	6.67	4.99±3.56	4.52	5.66±2.54	5.18
	Smina	7.08±3.47	7.43	4.88±3.69	4.38	7.25±4.15	7.41
	Gnina	6.67±3.65	6.51	4.57±3.56	3.74	6.72±3.95	6.06
	Glide	6.62±3.51	6.85	4.28±3.20	3.74	5.37±3.60	4.57
Traditional flexible docking methods	VinaFlex	8.81±2.94	9.06	8.80±2.98	9.07	7.36±3.33	7.07
	rDock	5.46±3.02	5.31	4.11±2.82	3.58	5.37±3.60	5.12
Deep learning-based docking methods	TankBind-pocket	2.63±1.88	1.97	2.37±2.38	1.63	3.74±1.98	3.42
	EDM-Dock	3.33±1.55	3.09	3.06±1.92	2.55	4.72±1.95	4.61
	KarmaDock Align	2.47±1.56	2.04	2.36±1.57	1.89	4.30±2.20	3.91
	DiffBindFR-Smina	2.96±2.32	2.10	2.30±2.28	1.58	4.92±4.12	4.02
	DiffBindFR-MDN	**2.32**±1.73	**1.76**	**1.87**±1.93	**1.25**	**3.64**±2.59	**2.74**

Table S6. (continued)

Method		DUDE27-HoloEns	
		RMSD Mean	RMSD Median
Traditional rigid receptor docking methods	Vina	6.04±3.37	5.88
	LinF9	5.70±3.16	5.47
	Smina	6.10±3.38	6.33
	Gnina	5.86±3.51	5.63
	Glide	5.72±3.41	5.38
Traditional flexible docking methods	VinaFlex	8.45±3.16	8.74
	rDock	5.00±3.31	4.44
Deep learning-based docking methods	TankBind-pocket	3.14±2.32	2.40
	EDM-Dock	3.64±2.00	3.29
	KarmaDock Align	**3.14**+2.36	2.37
	DiffBindFR-Smina	3.91±3.59	2.48
	DiffBindFR-MDN	3.28±3.31	**2.08**

$^{\alpha}$Best performance in bold for the lowest RMSD Mean and RMSD Medium.

Table S7. Performance of various methods on the 660 CASF2016 Apo-Holo pairs. $^{\alpha}$

Method		RMSD Mean	RMSD Median	PB-success rate
Traditional rigid receptor docking methods	Vina	5.94±3.66	5.62	0.195
	LinF9	5.60±3.47	5.30	0.186
	Smina	5.89±3.75	5.51	0.202
	Gnina	5.58±3.65	5.26	0.211
	Glide	5.19±3.18	5.01	0.132
Traditional flexible docking methods	VinaFlex	8.73±2.88	8.91	0.015
	rDock	5.17±2.82	4.86	0.148
Deep learning-based docking methods	TankBind-pocket	2.58±2.67	1.76	0.115
	EDM-Dock	3.67±12.5	2.73	0.053
	KarmaDock Align (release)	2.55±1.78	2.01	0.088
	KarmaDock Align (re-train)	2.47±1.61	2.00	0.109
	DiffBindFR-Smina	2.71±2.45	1.91	0.495
	DiffBindFR-MDN	**2.20**±2.15	**1.48**	**0.561**

$^{\alpha}$Best performance in bold. RMSD Mean and RMSD Medium, lowest; PB-success rate, highest. RMSD Mean and RMSD Median denote the average ± standard deviation and median of Ligand RMSD for top-1 generated ligand poses from each complex, respectively. KarmaDock Align (**release**) represents the released KarmaDock model trained on PDBbind general set, and KarmaDock Align (**re-train**) represents the KarmaDock model trained on PDBbind time-split training set.

Table S8. Performance of various methods on the 1100 CASF2016 Holo-Holo pairs.[α]

Method		RMSD Mean	RMSD Median	PB-success rate
Traditional rigid receptor docking methods	Vina	4.32±3.46	3.64	0.354
	LinF9	4.62±3.57	3.88	0.324
	Smina	4.28±3.52	3.48	0.365
	Gnina	3.96±3.37	2.95	0.385
	Glide	3.75±3.08	2.85	0.272
Traditional flexible docking methods	VinaFlex	8.84±3.03	9.16	0.027
	rDock	3.47±2.62	2.94	0.385
Deep learning-based docking methods	TankBind-pocket	2.24±2.18	1.59	0.128
	EDM-Dock	5.62±1.85	2.48	0.071
	KarmaDock Align (release)	2.26±1.58	1.78	0.132
	KarmaDock Align (re-train)	2.30±1.54	1.84	0.153
	DiffBindFR-Smina	2.11±2.14	1.36	0.609
	DiffBindFR-MDN	**1.67**±1.76	**1.12**	**0.682**

[α] Best performance in bold. RMSD Mean and RMSD Medium, lowest; PB-success rate, highest. RMSD Mean and RMSD Median denote the average ± standard deviation and median of Ligand RMSD for top-1 generated ligand poses from each complex, respectively. KarmaDock Align (**release**) represents the released KarmaDock model trained on PDBbind general set, and KarmaDock Align (**re-train**) represents the KarmaDock model trained on PDBbind time-split training set.

Table S9. Performance of various methods on DUDE27-HoloEns redocking pairs.[α]

Method		RMSD Mean	RMSD Median	PB-success rate
Traditional rigid receptor docking methods	Vina	2.98±3.31	1.22	0.565
	LinF9	3.37±3.42	1.80	0.540
	Smina	2.63±3.22	**1.10**	0.610
	Gnina	2.60±3.10	1.11	0.610
	Glide	3.14±3.32	1.88	0.530
Traditional flexible docking methods	VinaFlex	8.20±3.51	8.52	0.050
	rDock	2.67±2.75	1.42	0.560
Deep learning-based docking methods	TankBind-pocket	2.71±1.92	1.98	0.090
	EDM-Dock	3.44±2.21	2.98	0.090
	KarmaDock Align (re-train)	3.17±2.58	2.35	0.070
	DiffBindFR-Smina	**2.34**±2.93	1.18	**0.660**
	DiffBindFR-MDN	2.63±3.42	1.50	0.550

[α] Best performance in bold for the lowest RMSD Mean and RMSD Medium.

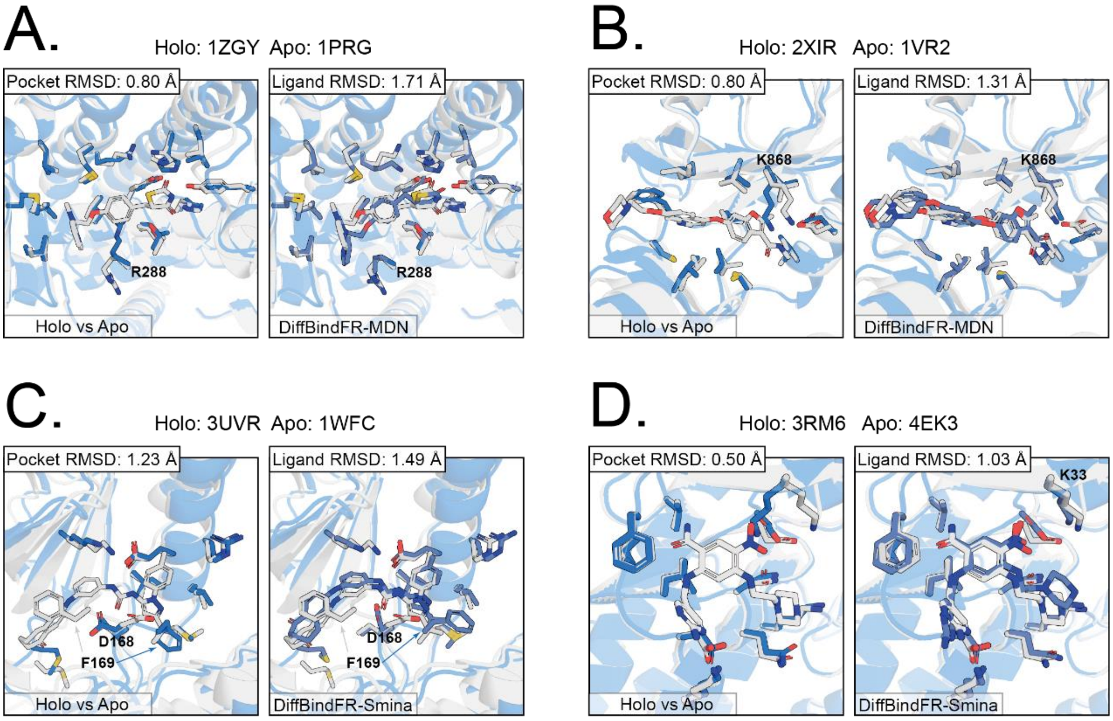

Fig. S11. The binding poses of four cases from ApoRef subset in CD test set. In all panels, Holo protein and ligand are shown in grey. Apo protein structure is shown in blue. DiffBindFR sampled ligand and pocket side chains are shown in sliver lake blue. Note that DiffBindFR sampled structure shares the backbone with Apo structure.

Table S10. Details about DUDE27-AF2 set.

Target	PDB & chain	UniProt ID	Backbone RMSD$^{\alpha}$	Flexible Side Chains for VinaFlex$^{\beta}$	Flexible residues
aces	1E66-A	P04058	0.41	ASP93, TRP105, GLU220, PHE351, TYR355, TRP453, HIS461	7
akt2	3D0E-A	P31751	0.88	LYS181, GLU200, MET229, TYR231, GLU279, MET282, THR292, ASP293, PHE294	9
bace1	3L5D-A	P56817	0.46	LEU91, ASP93, SER96, TYR132, THR133, GLN134, ASP289, THR292, THR293	9
hs90a	1UYG-A	P07900	0.78	ASN51, MET98, LEU107, PHE138, TRP162	5
tgfr1	3HMM-A	P36897	0.43	ILE211, LYS232, TYR282, HIS283, LEU340, ASP351	6
tryb1	2ZEC-A	Q15661	0.33	ASP218, SER219, GLN221, TRP244, GLU246	5
try1	2AYW-A	P00760	0.40	HIS63, LEU104, TYR154, GLN197, SER200, TRP216	6
thrb	1YPE-H	P00734	2.95	HIS406, TYR410, LEU459, TRP511, TRP590, PHE602	6
fabp4	2NNQ-A	P15090	0.67	PHE17, MET21, MET41, SER54, ILE105, ARG127, TYR129	7
ppard	2ZNP-A	Q03181	0.49	ARG248, LEU294, VAL298, LEU303, VAL312, LYS331, PHE332, HIS413, TYR43	9
pparg	2GTK-A	P37231	0.45	PHE292, HIS294, PHE310, GLN314, ARG316, SER317, HIS351, LEU358, PHE391, HIS477, TYR501	11
fa10	3KL6-A	P00742	0.58	THR318, TYR319, PHE396, GLN416, TRP439, GLU441, ILE451, TYR452	8
cdk2	1H00-A	P24941	0.83	ILE10, LYS33, PHE80, ASP86, LYS89, ASN132, LEU134, ASP145	8
met	3LQ8-A	P08581	1.09	LYS1110, GLU1127, MET1131, LEU1157, MET1211, ASP1222, PHE1223	7
mk10	2ZDT-A	P53779	0.91	LYS68, ILE70, LYS93,	7

				MET146, LEU148, MET149, ASN152	
rxra	1MV9-A	P19793	0.39	ILE268, GLN275, LEU309, ILE310, PHE313, ARG316, LEU326, ILE345, PHE346, HIS435, LEU436	11
mk14	2QD9-A	Q16539	1.32	VAL30, LYS53, ILE84, LEU104, THR106, LEU108, MET109, ASP112, LEU167	9
braf	3D4Q-A	P15056	1.38	ILE463, LYS483, LEU514, TRP531, PHE583, ASP594	6
vgfr2	2P2I-A	P35968	0.95	LYS868, GLU885, PHE918, CYS1024, LEU1035, ASP1046, PHE1047	7
gria2	3KGC-B	P19491	1.66	GLU423, TYR471, THR501, ARG506, LEU671, THR707, GLU726, MET729, TYR753	9
egfr	2RGP-A	P00533	1.04	LYS745, MET766, LEU777, THR790, MET793, LEU844, THR854, ASP855, PHE856	9
mapk2	3M2W-A	P49137	0.64	LEU70, LYS93, MET138, LEU141, ASP142, GLU190, LEU193, THR206, ASP207	9
ital	2ICA-A	P20701	0.89	ILE151, ILE260, ILE280, TYR282, ILE284, LYS312, LEU327, LYS330, ILE331	9
dpp4	2I78-B	P27487	0.34	GLU206, SER209, PHE357, ARG358, TYR547, SER630, ARG669	7
ptn1	2AZR-A	P18031	0.30	TYR46, ASP48, LYS120, ASP181, PHE182, CYS215, SER216, ILE219, ARG221, GLN262	10
igf1r	2OJ9-A	P08069	1.61	LEU1005, VAL1013, LYS1033, MET1079, MET1082, MET1142	6
ampc	1L2S-B	P00811	0.39	SER80, GLN136, ASN168, ARG220, TYR237	5

[a]The backbone RMSD is calculated between residues within 5 Å around crystal ligand in the Holo and matched ones in the AF2 predicted protein pocket.

[β]The flexible side chains are picked by expert experience through observing the binding-site aligned AF2 predicted and Holo crystal structures. Then, VinaFlex uses theses flexible side chains for flexible docking.

Table S11. The L-RMSD of various methods on DUDE27-AF2 test set.

Target	L-RMSD					
	IFD-MD (Holo) [α]	**IFD-MD (Docked)[β]**	**DiffBindFR-Smina**	**DiffBindFR-MDN**	**VinaFlex**	**rDock**
aces	0.68	6.19	0.59	0.65	0.75	5.63
akt2	0.57	10.82	1.16	1.79	9.12	5.53
bace1	1.20	4.96	4.61	3.29	8.32	4.99
hs90a	6.94	6.79	3.99	3.30	7.02	7.19
tgfr1	0.42	0.43	1.32	1.73	6.99	1.29
tryb1	5.22	5.19	1.25	1.24	9.48	2.00
try1	2.95	8.47	1.17	1.43	8.66	2.57
thrb	2.82	6.86	7.90	26.67	8.58	8.56
fabp4	0.96	7.14	5.44	5.57	4.37	4.80
ppard	1.42	0.85	3.49	2.12	1.42	4.03
pparg	1.84	1.49	1.17	1.31	3.31	3.45
fa10	1.26	1.36	1.62	0.95	10.32	0.82
cdk2	2.63	3.12	3.27	3.78	8	8.36
met	6.67	6.70	8.09	7.72	13.22	11.63
mk10	0.74	0.76	2.15	2.38	9.64	7.77
rxra	1.95	4.27	1.81	5.22	3.24	1.29
mk14	2.4	8.97	1.73	1.87	9.21	11.20
braf	1.38	5.75	1.4	1.28	5.02	1.58
vgfr2	1.38	7.85	8.3	2.94	10.83	10.53

gria2	1.74	3.71	10.96	2.07	5.65	6.08
egfr	2.19	10.12	5.20	1.96	9.68	11.21
mapk2	1.59	1.83	1.22	1.28	1.54	1.16
ital	1.63	6.86	6.78	4.24	10.87	11.07
dpp4	3.14	3.10	6.09	2.12	9.58	9.16
ptn1	0.53	1.24	1.91	1.98	1.78	0.64
igf1r	2.44	6.66	2.29	2.28	5.82	7.04
ampc	2.66	2.20	1.99	3.06	2.71	1.94
Median	1.74	5.19	2.15	2.12	8.00	5.53
SR[γ]	0.59	0.26	0.48	0.44	0.15	0.26
PB-SR[δ]	0.48	0.26	0.44	0.33	0.07	0.26

[α] IFD-MD (Holo) refers to utilize the ground-truth crystal ligand poses as the IFD-MD template poses.

[β] IFD-MD (Docked) refers to utilize the glide docked ligand poses as the IFD-MD template poses.

[γ] SR: The success rate of L-RMSD below 2 Å.

[δ] PB-SR: The PB-success rate.

Table S12. The sc-RMSD of various methods on DUDE27-AF2 test set.

Target	sc-RMSD for Flexible Pocket Side Chains[α]						
	Holo vs AF2	IFD-MD (Holo)	IFD-MD (Docked)	DiffBindFR-Smina	DiffBindFR-MDN	VinaFlex	rDock
aces	1.25	1.8	1.78	0.91	0.58	0.93	5.63
akt2	1.89	1.4	1.72	1.65	1.48	1.96	5.53
bace1	0.32	1.26	1.42	0.42	0.37	0.75	4.99
hs90a	2.12	2.08	2.11	1.54	1.59	2.23	7.19
tgfr1	0.98	1.42	1.35	1.2	1.57	0.87	1.29
tryb1	1.78	1.24	1.40	1.37	1.56	1.77	2.00
try1	1.48	1.54	1.35	0.94	1.33	2.67	2.57
thrb	3.44	3.21	3.31	3.44	4.95	4.75	8.56
fabp4	1.49	1.62	1.37	1.41	1.98	1.74	4.80
ppard	1.27	1.7	1.20	1.76	1.49	1.23	4.03
pparg	1.77	1.79	1.82	1.54	1.38	1.24	3.45
fa10	1.49	1.33	3.68	1.16	1.5	1.77	0.82
cdk2	1.60	1.46	2.32	2.33	2.84	1.85	8.36
met	5.11	5.03	5.38	5.34	5.32	5.69	11.63
mk10	1.44	1.1	1.20	1.35	1.49	2.19	7.77
rxra	1.18	1.46	1.61	1.68	2.42	1.70	1.29
mk14	1.48	1.54	1.57	1.38	1.4	1.92	11.20
braf	1.69	2.36	1.93	1.35	1.51	1.91	1.58
vgfr2	5.18	6.79	5.45	5.24	6.28	5.63	10.53

gria2	2.33	2.13	2.11	2.06	2.19	2.19	6.08
egfr	1.57	1.35	1.46	1.73	1.77	2.02	11.21
mapk2	1.23	1.31	1.58	1.03	1.49	1.66	1.16
ital	1.77	1.49	3.03	2.5	1.82	2.11	11.07
dpp4	1.68	1.84	1.65	1.49	1.95	3.23	9.16
ptn1	1.08	1.26	0.78	1.11	1.3	1.41	0.64
igf1r	1.70	1.75	1.69	1.59	1.52	2.00	7.04
ampc	1.87	2.50	1.92	1.43	1.27	2.14	1.94
Median	1.60	1.54	1.69	1.49	1.52	1.92	5.53
SR[β]	-	0.41	0.48	0.63	0.41	0.22	-

[α] “Flexible Pocket Side Chains” represents the “Flexible Side Chains for VinaFlex” recorded in Table S10.

[β] SR: The success rate of side chains refinement, defined by the sc-RMSD between the refined side chains and the corresponding side chains in Holo pocket is lower than the baseline (Holo vs AF2).

Table S12. (continued)

Target	sc-RMSD for Pocket Side Chains within 5 Å around the crystal ligand				
	Holo vs AF2	IFD-MD (Holo)	IFD-MD (Docked)	DiffBindFR-Smina	DiffBindFR-MDN
aces	1.12	1.45	1.37	0.90	0.93
akt2	2.36	2.80	2.26	2.19	2.10
bace1	0.99	1.47	1.86	0.83	0.89
hs90a	1.57	1.68	1.78	1.27	1.28
tgfr1	1.39	1.74	1.65	1.53	1.50
tryb1	1.24	1.19	1.40	0.97	0.98
try1	1.30	1.44	1.34	0.94	1.82
thrb	4.21	4.26	4.38	4.12	4.57
fabp4	1.42	1.57	1.86	1.53	1.57
ppard	1.33	1.53	1.28	1.45	1.31
pparg	1.53	1.37	1.62	1.47	1.33
fa10	1.38	1.43	2.65	1.43	1.55
cdk2	1.49	1.58	2.08	2.06	2.08
met	3.41	3.53	3.64	3.52	3.43
mk10	1.22	1.37	1.50	1.34	1.36
rxra	1.07	1.37	1.42	1.35	1.82
mk14	1.75	2.23	1.89	1.84	1.52
braf	2.22	2.63	2.80	2.12	2.33
vgfr2	3.15	4.00	3.45	3.19	3.70
gria2	2.09	2.09	2.01	1.96	2.03
egfr	2.49	2.20	3.38	2.57	2.61
mapk2	1.15	1.54	1.55	1.12	1.48
ital	1.54	1.35	3.11	1.71	1.79
dpp4	1.10	1.48	1.68	1.01	1.25
ptn1	0.99	1.09	0.71	0.97	1.12

igf1r	3.55	3.19	3.49	3.53	3.44
ampc	1.11	1.85	1.57	0.98	0.92
Median	1.42	1.57	1.86	1.47	1.55
SR	-	0.22	0.19	0.56	0.41

10.The discussion about the redocking success rate

The reported redocking performance of AutoDock Vina and Glide on the time-split test set is considerably lower compared to other literatures. We believe that there are two main reasons why conventional methods did not achieve a high redocking success rate (>80%) conducted in our work:

10.1 Dataset Differences

Currently, the most commonly used dataset in the docking field is PDBbind 2020. There are generally two ways to divide the PDBbind 2020 dataset: the MLSF-split and the PDBbind time-split. The MLSF-split methods designate 90% of the PDBbind general set as the training set, 10% of the PDBbind general set as the validation set, and use CASF2016 as the test set, a typical test set reporting high docking success rate. On the other hand, the time-split methods use 363 complex structures from the PDBbind 2020 dataset, uploaded after 2019, as the test set. After excluding ligands present in the test set, the remaining 16,739 structures are used for training, and 968 structures are used for validation.

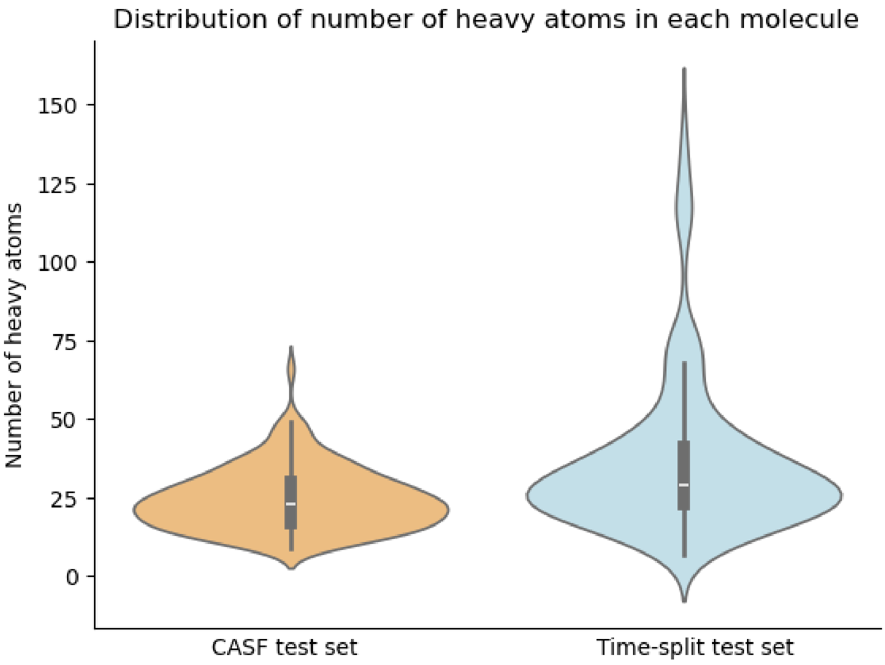

Fig. S12. Distribution of number of ligand-containing heavy atoms in CASF2016 test set and time-split test set.

Compared to the CASF2016 test set, the PDBbind time-split test set contains ligands with a higher number of heavy atoms (Fig. S12), even with about 15% of the ligands being peptides[34]. This composition results in traditional methods performing more poorly on the PDBbind time-split test set.

Additionally, compared to the CASF2016 test set, the PDBbind time-split test set contains a higher number of structures with poor resolution (Fig. S13), which can also make re-dock harder.

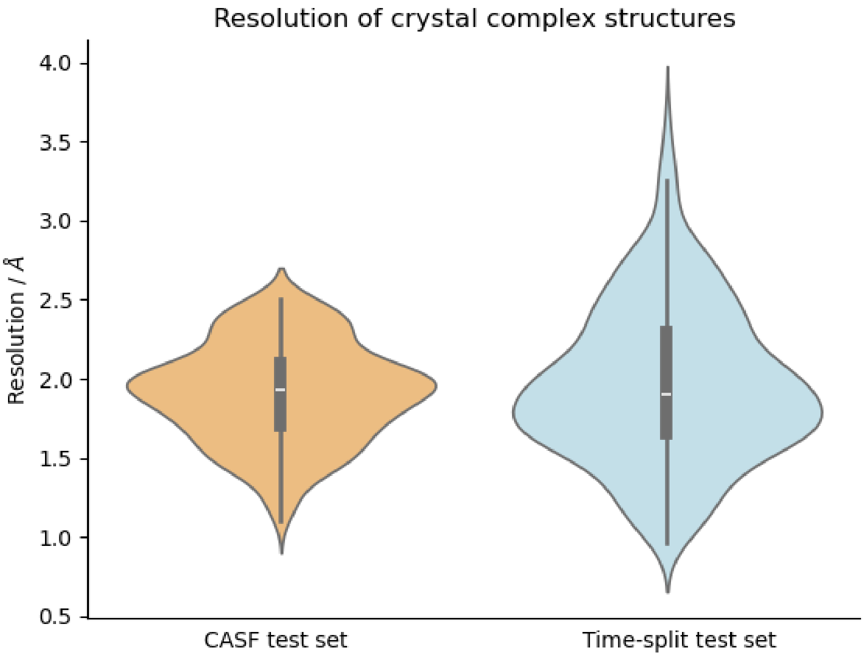

Fig. S13. Resolution of crystal complex structures in CASF2016 test set and time-split test set.

10.2 Exhaustive Sampling

In order to evaluate the real-world applicability of docking methods, we did not perform exhaustive sampling for each docking program. The docking power assessment of the CASF2016 test set provided decoy poses[33] generated by re-docking using three molecular docking programs, including GOLD (version 5.2, Cambridge Crystallographic Data Center), Surflex implemented in the SYBYL software (version 8.1, CERTARA Inc.), and the molecular docking module implemented in the MOE software (version 2015, Chemical Computing Group). GOLD was used to generate 400 binding poses, Surflex generated 300 binding poses, and MOE also generated 300 binding poses. Residues within 10 Å from the native ligand were considered to form the binding pocket. Through pose clustering, up to 100 binding poses were selected as representatives and used in the docking power set. The representatives' decoys from the CASF2016 docking power set, using Vina as its scoring function, selected the top1 ligand binding pose achieving an 81.2% success rate (Table. S13). CASF2016 pays more attention to the docking power of scoring functions as the decoy poses are well prepared. Indeed, a practical docking process involving pose sampling and scoring, both of them decide the docking success rate. Therefore, we further used Glide and Vina to perform “sampling-and-scoring” docking, and carefully prepared proteins and ligands following the official guidelines (consistent with those used in our paper), considering residues within 10 Å from the native ligand to form the binding pocket. Subsequently, 40 rounds of independent sampling were performed on the complexes in this test set, with the top-scoring ligand pose used to calculate L-RMSD. To prevent confusion with later results, we refer to the Vina method here as Vina buffer. The results showed that Glide and Vina buffer achieved docking success rates of 59.6% and 63.2% (Table. S13), respectively, on the CASF2016 test set. Furthermore, to prove that our baseline method of defining the pocket using residues within 10 Å of the ligand's center coordinate does not affect docking performances, we also tested Vina docking results using the same pocket definition as in our paper (Vina box center). The results revealed that Vina buffer and Vina box center achieved the same success rate (Table. S13), with a similar distribution of L-RMSD for the top1 ligand pose (Fig. S14). This indicates that achieving around an 80% success rate in re-docking tasks requires significant computational efforts and the use of multiple docking tools for exhaustive sampling, which is clearly impractical for screening large compound libraries.

Table S13. The L-RMSD of various methods on DUDE27-AF2 test set.

Strategy name	Sampling method	Scoring function	Pocket definition	Sampling turns	Success rate (L-RMSD < 2 A)
Vina center box	Vina	Vina	Residues within 24 Å cubic at native ligand coordinate center	40	63.2%
Vina buffer	Vina	Vina	Residues within 10 Å from the native ligand	40	63.2%
Glide	Glide SP	Glide SP	Residues within 10 Å from the native ligand	40	59.6%
CASF2016 Decoys	GOLD, Surflex, MOE	Vina	Residues within 10 Å from the native ligand	1000	81.2%

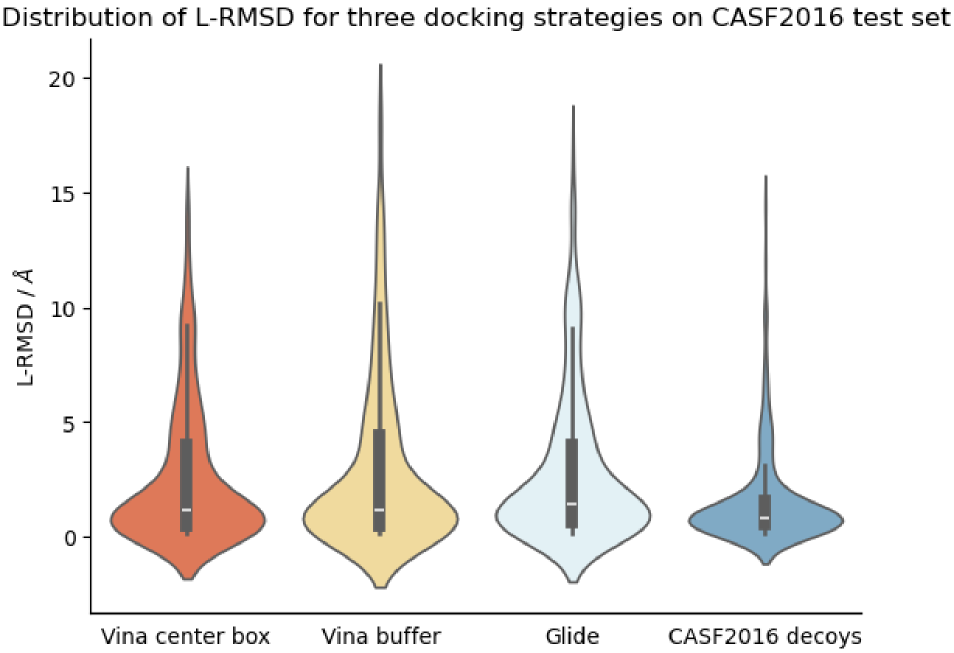

Fig. S14. L-RMSD distribution of selected top1 ligand pose using different conventional docking methods.

References

1. Zhang J., Li H., Zhao X., *et al.* Holo protein conformation generation from Apo structures by ligand binding site refinement. *Journal of Chemical Information and Modeling*, 2022, 62: 5806-5820.

2. Aggarwal R., Gupta A. and Priyakumar U. Apobind: a dataset of ligand unbound protein conformations for machine learning applications in de novo drug design. *arXiv* 2108.09926, 2021.

3. Feidakis C. P., Krivak R., Hoksza D., *et al.* AHoJ: rapid, tailored search and retrieval of apo and holo protein structures for user-defined ligands. *Bioinformatics*, 2022, 38: 5452-5453.

4. Zhang Y., Vass M., Shi D., *et al.* Benchmarking refined and unrefined Alphafold2 structures for hit discovery. *Journal of Chemical Information and Modeling*, 2023, 63: 1656-1667.

5. Miller E. B., Murphy R. B., Sindhikara D., *et al.* Reliable and accurate solution to the induced fit docking problem for protein-ligand binding. *Journal of Chemical Theory and Computation*, 2021, 17: 2630-2639.

6. Mysinger M. M., Carchia M., Irwin J. J., *et al.* Directory of useful decoys, enhanced (DUD-E): better ligands and decoys for better benchmarking. *Journal of Medicinal Chemistry*, 2012, 55: 6582-6594.

7. Schoning-Stierand K., Diedrich K., Ehrt C., *et al.* ProteinsPlus: a comprehensive collection of web-based molecular modeling tools. *Nucleic Acids Research*, 2022, 50: W611-W615.

8. Bietz S. and Rarey M. SIENA: efficient compilation of selective protein binding site ensembles. *Journal of Chemical Information and Modeling*, 2016, 56: 248-259.

9. Karelina M., Noh J. J. and Dror R. O. How accurately can one predict drug binding modes using AlphaFold models? *eLife*, 2023, 12: RP89386.

10. Jumper J., Evans R., Pritzel A., *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature*, 2021, 596: 583-589.

11. Hornak V., Abel R., Okur A., *et al.* Comparison of multiple Amber force fields and development of improved protein backbone parameters. *Proteins: Structure, Function, and Bioinformatics*, 2006, 65: 712-725.

12. Schrödinger LLC. Schrödinger Release 2021-2. https://www.schrodinger.com.

13. Banks J. L., Beard H. S., Cao Y., *et al.* Integrated modeling program, applied chemical theory (IMPACT). *Journal of Computational Chemistry*, 2005, 26: 1752-1780.

14. Sondergaard C. R., Olsson M. H., Rostkowski M., *et al.* Improved treatment of ligands and coupling effects in empirical calculation and rationalization of pKa values. *Journal of Chemical Theory and Computation*, 2011, 7: 2284-2295.

15. Olsson M. H., Sondergaard C. R., Rostkowski M., *et al.* PROPKA3: consistent treatment of internal and surface residues in empirical pKa predictions. *Journal of Chemical Theory and Computation*, 2011, 7: 525-537.

16. Landrum G. RDKit, 2023-03-10. https://github.com/rdkit/rdkit.

17. Eberhardt J., Santos-Martins D., Tillack A. F., *et al.* AutoDock Vina 1.2.0: new docking methods, expanded force field, and python bindings. *Journal of Chemical*

Information and Modeling, 2021, 61: 3891-3898.

18. Forli Lab. Meeko, 2023-07-29. https://github.com/forlilab/Meeko.

19. Koes D. R., Baumgartner M. P. and Camacho C. J. Lessons learned in empirical scoring with smina from the CSAR 2011 benchmarking exercise. *Journal of Chemical Information and Modeling*, 2013, 53: 1893-1904.

20. Yang C. and Zhang Y. Lin_F9: a linear empirical scoring function for protein-ligand docking. *Journal of Chemical Information and Modeling*, 2021, 61: 4630-4644.

21. Jones G., Willett P., Glen R. C.*, et al.* Development and validation of a genetic algorithm for flexible docking. *Journal of Molecular Biology*, 1997, 267: 727-748.

22. Buttenschoen M., Morris G. M. and Deane C. M. PoseBusters: AI-based docking methods fail to generate physically valid poses or generalise to novel sequences. *Chemical Science*, 2024, 15: 3130-3139.

23. Ravindranath P. A., Forli S., Goodsell D. S.*, et al.* AutoDockFR: advances in protein-ligand docking with explicitly specified binding site flexibility. *PLoS Computational Biology*, 2015, 11: e1004586.

24. Ruiz-Carmona S., Alvarez-Garcia D., Foloppe N.*, et al.* rDock: a fast, versatile and open source program for docking ligands to proteins and nucleic acids. *PLoS Computational Biology*, 2014, 10: e1003571.

25. Friesner R. A., Banks J. L., Murphy R. B.*, et al.* Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. *Journal of Medicinal Chemistry*, 2004, 47: 1739–1749.

26. Lu W., Wu Q., Zhang J.*, et al.* Tankbind: trigonometry-aware neural networks for drug-protein binding structure prediction, in *Advances in Neural Information Processing Systems*, 2022.

27. Masters M. R., Mahmoud A. H., Wei Y.*, et al.* Deep learning model for efficient protein-ligand docking with implicit side-chain flexibility. *Journal of Chemical Information and Modeling*, 2023, 63: 1695-1707.

28. Zhang X. J., Zhang O., Shen C.*, et al.* Efficient and accurate large library ligand docking with KarmaDock. *Nature Computational Science*, 2023, 3: 789-804.

29. Corso G., Jing B., Barzilay R.*, et al.* DiffDock: diffusion steps, twists, and turns for molecular docking, in *International Conference on Learning Representations*, 2023.

30. Jing B., Eismann S., Suriana P.*, et al.* Learning from protein structure with geometric vector perceptrons, in *International Conference on Learning Representations*, 2021.

31. Dwivedi V. P. and Bresson X. A generalization of transformer networks to graphs. *arXiv* 2012.09699, 2020.

32. Méndez-Lucio O., Ahmad M., del Rio-Chanona E. A.*, et al.* A geometric deep learning approach to predict binding conformations of bioactive molecules. *Nature Machine Intelligence*, 2021, 3: 1033-1039.

33. Su M., Yang Q., Du Y.*, et al.* Comparative assessment of scoring functions: the CASF-2016 update. *Journal of Chemical Information and Modeling*, 2019, 59: 895-913.

34. Masters M., Mahmoud A. H. and Lill M. A. PocketNet: ligand-guided pocket prediction for blind docking., in *ICLR 2023-Machine Learning for Drug Discovery*

workshop, 2023.