Electronic Supplementary Material (ESI) for Chemical Science.
This journal is © The Royal Society of Chemistry 2023

Supporting Information

A generalized protein-ligand scoring framework with balanced scoring, docking, ranking and screening powers

Chao Shen[1,2,3,4], Xujun Zhang[1], Chang-Yu Hsieh[1], Yafeng Deng[4], Dong Wang[1], Lei Xu[5], Jian Wu[3], Dan Li[1], Yu Kang[1,*], Tingjun Hou[1,2,*], Peichen Pan[1,*]

[1]Innovation Institute for Artificial Intelligence in Medicine of Zhejiang University, College of Pharmaceutical Sciences, Zhejiang University, Hangzhou 310058, Zhejiang, China

[2]State Key Lab of CAD&CG, Zhejiang University, Hangzhou 310058, Zhejiang, China

[3]School of public health, Zhejiang University, Hangzhou 310058, Zhejiang, China

[4]CarbonSilicon AI Technology Co., Ltd, Hangzhou 310018, Zhejiang, China

[5]Institute of Bioinformatics and Medical Engineering, School of Electrical and Information Engineering, Jiangsu University of Technology, Changzhou 213001, China

Corresponding authors

Yu Kang

E-mail: yukang@zju.edu.cn

Tingjun Hou

E-mail: tingjunhou@zju.edu.cn

Peichen Pan

E-mail: panpeichen@zju.edu.cn

Table S1. Input node and edge features for ligand graph representation.

Features	Size	Description
Nodes		
atom_type_one_hot	17	One hot encoding for atom type ([“C”, “N”, “O”, “S”, “F”, “P”, “Cl”, “Br”, “I”, “B”, “Si”, “Fe”, “Zn”, “Cu”, “Mn”, “Mo”, “other”])
atom_degree_one_hot	7	One hot encoding for atom degree ([0, 1, 2, 3, 4, 5, 6])
atom_formal charge	1	Formal charge
atom_num_radical electrons	1	Number of radical electrons
atom_hybridization_one_hot	6	One hot encoding for atom hybridization ([“sp”, “sp2”, “sp3”, “sp3d”, “sp3d2”, “other”])
atom_is_aromatic	1	Whether the atom is aromatic
atom_total_num_H_one_hot	5	One hot encoding for total number of Hs one the atom ([0, 1, 2, 3, 4])
atom_chirality_one_hot	3	One hot encoding for chirality of an atom ([“R”, “S”, “other”])
Edges		
bond_type_one_hot	4	One hot encoding for bond type ([“SINGLE”, “DOUBLE”, “TRIPLE”, “AROMATIC”])
bond_is_conjugated	1	Whether the bond is conjugated
bond_is_in_ring	1	Whether the bond is in a ring
bond_stereo_one_hot	4	One hot encoding for the stereo configuration of a bond (["STEREONONE", "STEREOANY", "STEREOZ", "STEREOE"])

Table S2. Input node and edge features for protein pocket graph representation.

Features	Size	Description
Nodes		
residue_type_one_hot	32	One hot encoding for residue type ([“GLY”, “ALA”, “VAL”, “LEU”, “ILE”, “PRO”, “PHE”, “TYR”, “TRP”, “SER”, “THR”, “CYS”, “MET”, “ASN”, “GLN”, “ASP”, “GLU”, “LYS”, “ARG”, “HIS”, “MSE”, “CSO”, “PTR”, “TPO”, “KCX”, “CSD”, “SEP”, “MLY”, “PCA”, “LLP”, “metal”, “other”]).
residue_self_distance	5	The maximum and minimum values of the scaled distance (multiplied by 0.1) within any atom in a residue, and the scaled distance (multiplied by 0.1) between the atoms named as CA and O, the distance (multiplied by 0.1) between the atoms named as O and N, and the distance (multiplied by 0.1) between the atoms named as C and N.
residue_dihedral _angle	4	The scaled dihedral angles (multiplied by 0.01), including *phi*, *psi*, *omega* and *chi1*.
Edges		
residue_is_connected	1	Whether two residues are covalently connected.
residue_CA_distance	1	The scaled distance (multiplied by 0.1) between the CA atoms of two residues.
residue_center_distance	1	The scaled distance (multiplied by 0.1) between the center of two residues.
residue_maximum_distance	2	The maximum and minimum values of the scaled distance (multiplied by 0.1) between two residues.

Table S3. The hyperparameter setting of the model.

Hyperparameters	Setting	
	GT	GatedGCN
Hidden dimension of GT/GatedGCN layer (d)	128	
Number of attention heads (H)	4	/
Number of GT/GatedGCN layers (L)	6	
Hidden dimension of mixture density network (d_m)	128	
Number of Gaussians (N_g)	10	
Dropout rate	0.15	
Learning rate	10^{-3}	
Weight decay	10^{-5}	
Maximum number of epochs	5000	
Batch size	64	
Patience of early stopping	70	

Table S4. The basic information of the LIT-PCBA dataset employed in this study

Target	Target name	Actives/Inactives		PDB entry
		Before Docking	After Docking	
ADRB2	Beta2 adrenergic receptor	17/312483	16/299194	4LDO
ALDH1	Aldehyde dihydrogenase 1	7168/137965	7162/137694	5L2M
ESR_ago	Estrogen receptor α	13/5583	10/5341	2P15
ESR_antago	Estrogen receptor α	102/4948	101/4910	2IOK
FEN1	FLAP Endonuclease	369/355402	368/354921	5FV7
GBA	Glucocerebrosidrase	166/296052	166/295663	2V3D
IDH1	Isocitrate dihydrogenase	39/362049	39/361378	4UMX
KAT2A	Histone acetyltransferase KAT2A	194/348548	193/348099	5H86
MAPK1	Mitogen-activated protein kinase 1	308/62629	308/62525	4ZZN
MTORC1	Mechanistic target of rapamycin	97/32972	97/32966	4DRI
OPRK1	Kappa opioid receptor	24/269816	23/269418	6B73
PKM2	Pyruvate kinase muscle isoform 2	546/245523	544/244137	4JPG
PPARG	Peroxisome proliferator-activated receptor γ	27/5211	27/5198	5Y2T
TP53	Cellular tumor antigen p53	79/4168	79/4146	3ZME
VDR	Vitamin D receptor	884/355388	847/345609	3A2I

Table S5. The training sets for the SFs summarized in Table 1.

Method	Training set[a]
KORP-PL	PDBbind-v2016 general set (N = 12910)
K_{DEEP}	PDBbind-v2016 refined set (N = 3772)
AKScore	PDBbind-v2016 refined set (N = 3772)
$\Delta_{Vina}RF_{20}$	PDBbind-v2014 refined set + natives poses in the CSAR decoy set + weak-binding crystal structures in PDBbind-v2014 general set (N = 3336) + CSAR decoy set (N = 3322)
Δ_{Vina}XGB	structures in PDBbind-v2016 refined set released before 2015 (N = 3565 without water + 3257 with receptor-bound water) + decoys based on CSAR and PDBbind (N = 7584)
Δ_{Lin_F9}XGB	PDBbind-v2016 refined set after filtering (N = 6816) + weak and strong binders from PDBbind-v2018 general set (N = 1556 + 510) + strong binders obtained by flexible-redocking (N = 235) + decoys generated through docking (N = 7111 + 5715)
OnionNet-SFCT+Vina	PDBbind-v2018 general set (N = 12906) + Docking poses generated based on PDBbind-v2018 general set (N = 10208×90)
AEScore	PDBbind-v2016 refined set (N = 3377)
Δ-AEScore	PDBbind-v2016 refined set (N = 3377)
PIGNet	PDBbind-v2019 refined set (N = 4514) + decoys generated based on PDBbind-v2016 (N = 292518 + 831885 + 527682)
DeepDock	PDBbind-v2019 general set (N = 16367)
RTMScore	PDBbind-v2020 general set (N = 19149)
Models in this study	PDBbind-v2020 general set (N = 19149)

[a]for all the methods, the structures in the training set that are overlapped with the ones in test set shall have been eliminated.

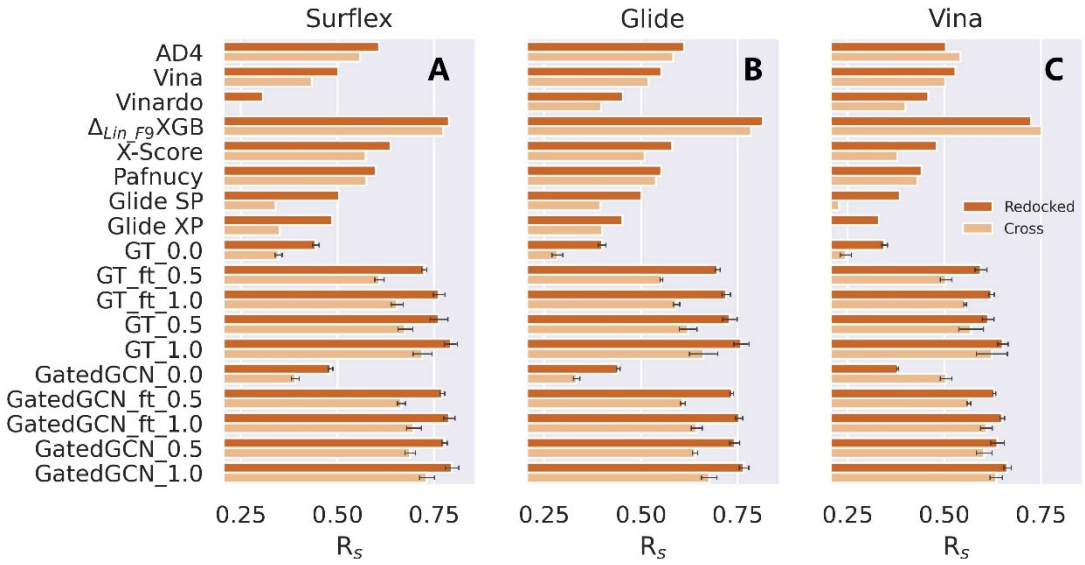

Figure S1. Scoring powers of scoring functions on PDBbind-CrossDocked-Core set indicated by Spearman correlation coefficient (R_s), where the poses are generated by (A) Surflex-Dock, (B) Glide SP and (C) AutoDock Vina, respectively.