Supporting Information

Ligand binding affinity prediction with fusion of graph neural networks and 3D structure-based complex graph

Lina Dong, [a] Shuai Shi, [c] Xiaoyang Qu, [a] Ding Luo [a] and Binju Wang*[a, b]

[a]State Key Laboratory of Physical Chemistry of Solid Surfaces and Fujian Provincial Key Laboratory of Theoretical and Computational Chemistry, iChEM, College of Chemistry and Chemical Engineering, Xiamen University, Xiamen, China, 361005
[b]Innovation Laboratory for Sciences and Technologies of Energy Materials of Fujian Province (IKKEM), Xiamen, China, 361005
[c] Department of Algorithm, TuringQ Co., Ltd., Shanghai, China, 200240
*Correspondence to: wangbinju2018@xmu.edu.cn

Table S1. The node features and edge features employed in the protein and ligand graph construction.

Type	Level	Attributes name	Descriptions	Length
Node Features	2D	Atom type	Encoding for atom type (['B', 'C', 'N', 'O', 'P', 'S', 'Se', 'halogen', 'metal'])	9×2 (protein/ligand)
		Atom properties	['hyb','heavyvalence','heterovalence','partialcharge'] is used	4×2
		Hytrophobic	Whether the atom is hydrophobic	1×2
		Aromatic	Whether the atom has aromaticity	1×2
		Hydrogen bond	['acceptor', 'donor'] is used	2×2
		Ring	Whether the atom on the ring	1×2
Edge Features	3D	Distance	The scaled Euclidean distance (multiplied by 0.1) between the connected atoms in 3D space.	1
		Distance statistics	The max, sum and mean values of scaled distances (multiplied by 0.1) between atoms i, k in 3D space	3
		Angle statistics	The max, sum and mean values of scaled (multiplied by 0.01) angle between atoms i, j, k in 3D space	3
		Area statistics	The max, sum and mean values of areas between atoms i, j, k in 3D space	3
		RBF-distance	Discretize the distance with 15 as the resolution	15
Total				57

Table S2. Model parameters for FGNN.

Model name	Parameters

SignNet	in_channel=256
	hidden_channel=256
	out_channel=128
	edge_dim=10

Attentive_FP	in_channel=36+128(node_dim+SignNet out_dim)
	hidden_channel=256
	out_channel=128
	edge_dim=10
	num_layers=3
	num_timesteps=3

Regression_layer	in_channel=128
	hidden_channel_1=1024
	hidden_channel_2=512
	out_channel=1
	dropout=0.1

Table S3. Train parameters for FGNN.

Type	Prameters

lr_scheduler setting	mode=min
	factor=0.5
	cooldown=30
	min_lr=1e-6

kfold setting	kfold=5
	shuffle=True

dataloader setting:	batch_size=64

other setting	epoch=300
	lr=0.01

Table S4. Performance of individual models and fusion models on PDBbind2016 crystal structures. The training set is PDBbind 2016 general and refined set (12906) in Table 1. The test set consists of 285 crystal structures tested for scoring power in CASF-2016.

Model	Training set			Test set		
	Rp	Rs	RMSE	Rp	Rs	RMSE
GIN	0.989	0.989	0.33	0.847	0.842	1.22
GIN+3D (GINE)	0.987	0.992	0.32	0.838	0.828	1.22
SignNet	0.990	0.990	0.37	0.536	0.523	1.85
SignNet+3D	0.912	0.888	0.89	0.764	0.746	1.46
Attentive_FP	0.987	0.985	0.37	0.819	0.800	1.30
Attentive_FP+3D	0.992	0.992	0.27	0.850	0.839	1.19
FGNN1 (Fusion of GIN and Attentive_FP+3D)	0.992	0.992	0.26	0.854	0.846	1.17
FGNN2 (Fusion of GIN+3D and Attentive_FP+3D)	0.992	0.992	0.27	0.869	0.865	1.13
FGNN3 (Fusion of SignNet+3D and Attentive_FP+3D)	0.993	0.993	0.26	0.873	0.867	1.14

Part S1. Results of data augmentation.

Besides crystal structures from PDBbind2016[1] general and refined set, we selected comparable number of rigid decoys (12000) from CSAR-decoys set[2] as negative samples for training. The labels of these decoys are defined in the same way as Section 2.1. The results are as follows (Table S2). For the convenience of comparison, we also list the results without data augmentation below, and those with data expansion are identified by DA. In addition to FGNN3, the results of data augmentation and retraining of other models have improved compared with the baselines. However, data augmentation has little effect on scoring power of SignNet[3] and Attentive_FP[4]. Data augmentation has a negative impact on the scoring power of FGNN3, possibly due to the pseudo label setting rules, data quality and model capacity. How to further improve the performance of the large parameter capacity model (such as FGNN3) through data will also be the direction of our future efforts.

Table S5. Impact of data augmentation on scoring power.

	Training set			Test set		
	Rp	Rs	RMSE	Rp	Rs	RMSE
GIN+3D	0.987	0.992	0.32	0.838	0.828	1.22
GIN+3D+DA	0.996	0.996	0.20	0.850	0.843	1.19
SignNet+3D	0.912	0.888	0.89	0.764	0.746	1.46
SignNet+3D+DA	0.990	0.989	0.33	0.765	0.771	1.44
Attentive_FP+3D	0.992	0.992	0.27	0.850	0.839	1.19
Attentive_FP+3D+DA	0.996	0.996	0.21	0.855	0.840	1.18
FGNN1 (GIN+Attentive_FP+3D)	0.992	0.992	0.26	0.854	0.846	1.17
FGNN1 (GIN+Attentive_FP+3D)+DA	0.995	0.995	0.21	0.867	0.860	1.13
FGNN2 (GIN+3D+Attentive_FP+3D)	0.992	0.992	0.27	0.869	0.865	1.13
FGNN2 (GIN+3D+Attentive_FP+3D)+DA	0.996	0.996	0.20	0.871	0.860	1.12
FGNN3 (SignNet+Attentive_FP+3D)	0.993	0.993	0.26	0.873	0.867	1.14
FGNN3 (SignNet+Attentive_FP+3D)+DA	0.996	0.996	0.21	0.818	0.803	1.30

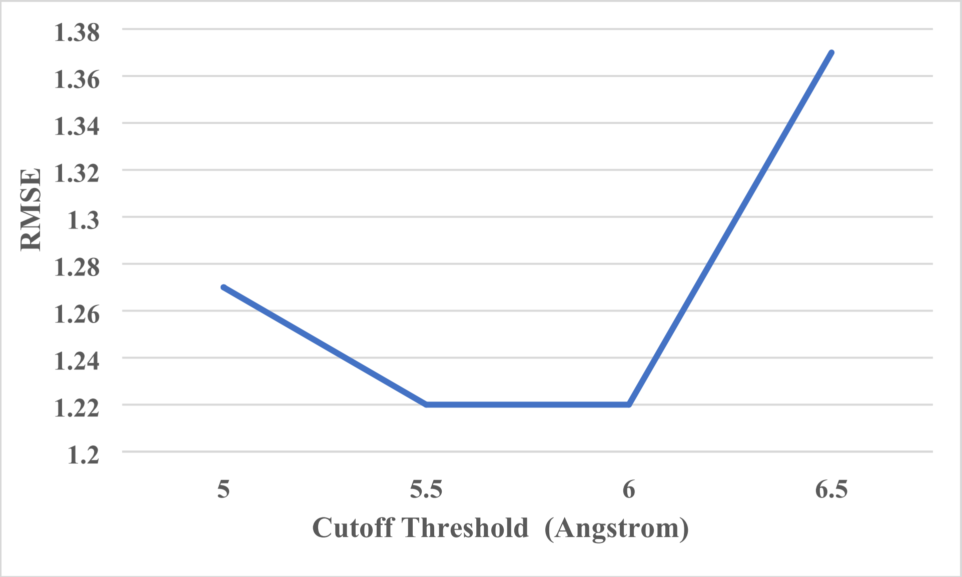

Fig.S1 Setting of the cutoff threshold. The RMSE of 5.5 Å is equal to 6 Å. Considering the computing resources, the threshold value set in our subsequent experiments is 5.5 Å.

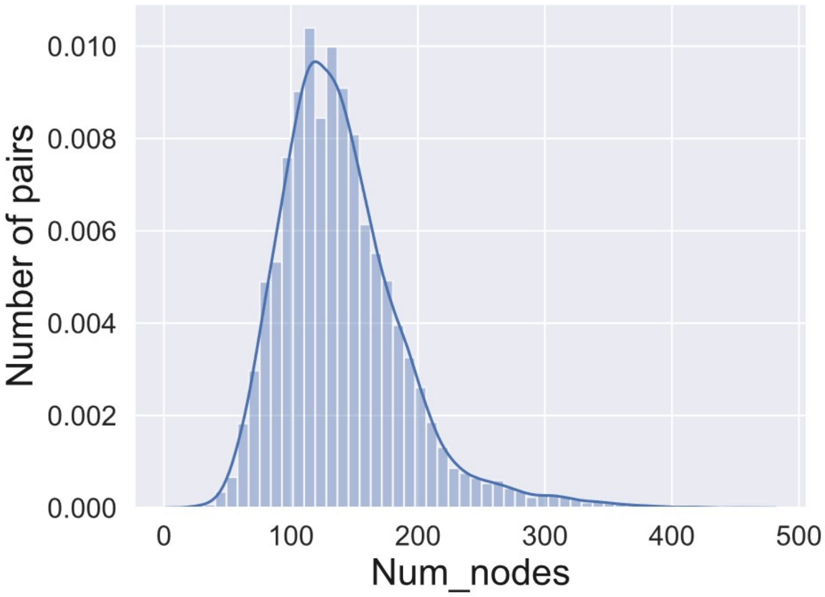

Fig.S2 Statistics of the number of nodes in composite graphs.

REFERENCES

(1) Wang, R.; Fang, X.; Lu, Y.; Wang, S., The PDBbind database: collection of binding affinities for protein-ligand complexes with known three-dimensional structures. *J Med Chem* **2004**, 47, 2977-80.
(2) Huang, S. Y.; Zou, X., Scoring and lessons learned with the CSAR benchmark using an improved iterative knowledge-based scoring function. *J Chem Inf Model* **2011**, 51, 2097-106.
(3) Lim, D.; Robinson, J.; Zhao, L.; Smidt, T.; Sra, S.; Maron, H.; Jegelka, S., Sign and Basis Invariant Networks for Spectral Graph Representation Learning. *arXiv preprint arXiv:2202.13013* **2022**.
(4) Xiong, Z.; Wang, D.; Liu, X.; Zhong, F.; Wan, X.; Li, X.; Li, Z.; Luo, X.; Chen, K.; Jiang, H.; Zheng, M., Pushing the Boundaries of Molecular Representation for Drug Discovery with the Graph Attention Mechanism. *J Med Chem* **2020**, 63, 8749-8760.