iScience, Volume *26*

Supplemental information

Protein–ligand binding affinity prediction

with edge awareness and supervised attention

Yuliang Gu, Xiangzhou Zhang, Anqi Xu, Weiqi Chen, Kang Liu, Lijuan Wu, Shenglong Mo, Yong Hu, Mei Liu, and Qichao Luo

Supplemental information

Protein–Ligand Binding Affinity Prediction with Edge Awareness and Supervised Attention

Authors:

Yuliang Gu[1,2,5], Xiangzhou Zhang[2,4,5], Anqi Xu[1,5], Weiqi Chen[2,4], Kang Liu[2], Lijuan Wu[2], Shenglong Mo[2], Yong Hu[2,4*], Mei Liu[3*], Qichao Luo[1,2,6*]

Affiliation of the authors:

[1] *Department of Pharmacology, School of basic medicine, Anhui Medical University, Hefei, Anhui, 230022, China.*

[2] *Big Data Decision Institute, Jinan University, Guangzhou, Guangdong, 510632, China.*

[3] *Division of Medical Informatics, Department of Internal Medicine, University of Kansas Medical Center, Kansas City, KS, 66160, USA*

[4] *The First Affiliated Hospital, Jinan University, Guangzhou, Guangdong, 510632, China.*

[5] These authors contributed equally

[6] Lead contact

[*] Correspondence: yonghu@jnu.edu.cn (H.Y.), meiliu@kumc.edu (L.M.), luoqichao@ahmu.edu.cn (L.Q.)

Supplemental Figures

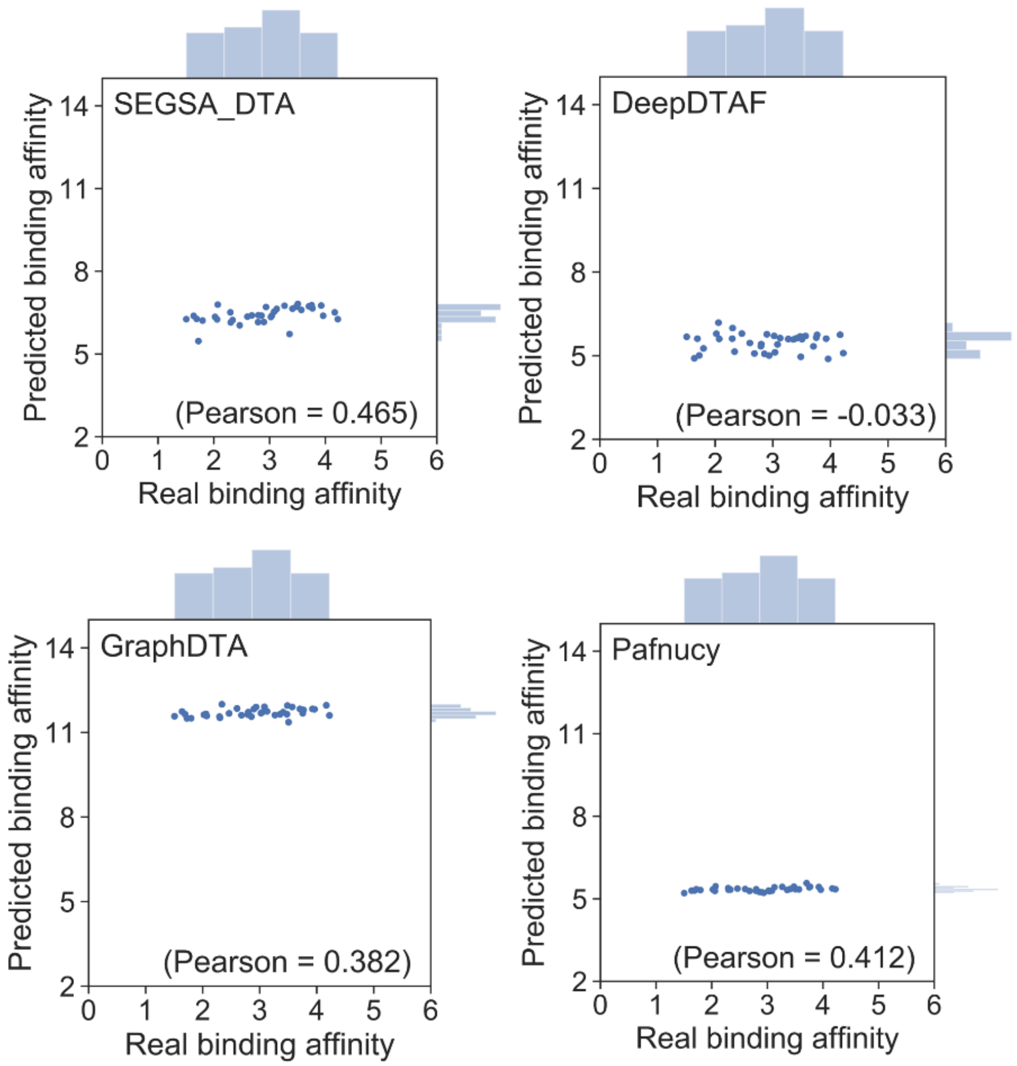

Figure S1. Distributions of predicted binding affinities on the Mpro_37 data set. Related to Table 1 and Figure 2A.

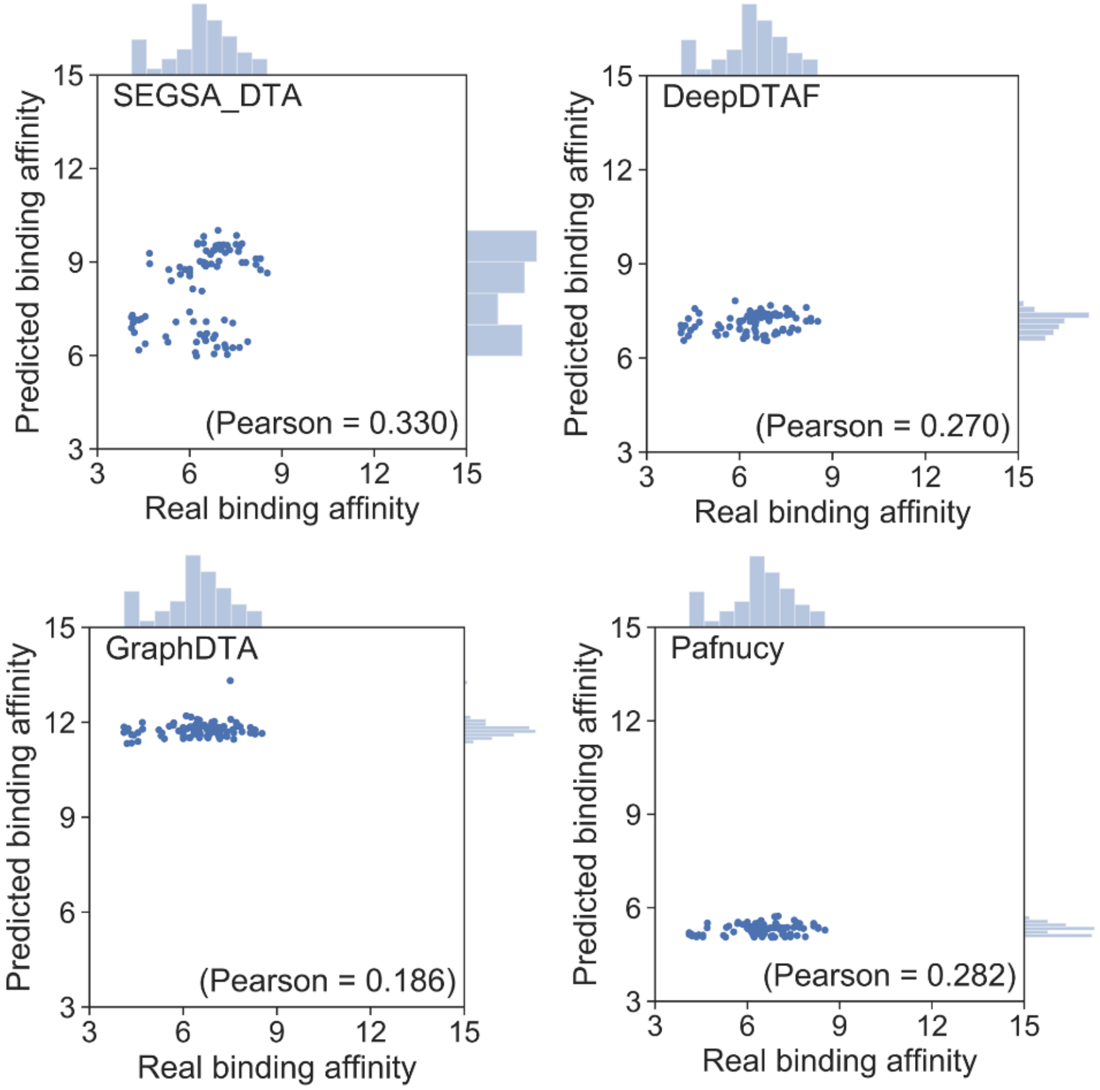

Figure S2. Distributions of predicted binding affinities on the PIM1_89 data set. Related to Table 1 and Figure 2A.

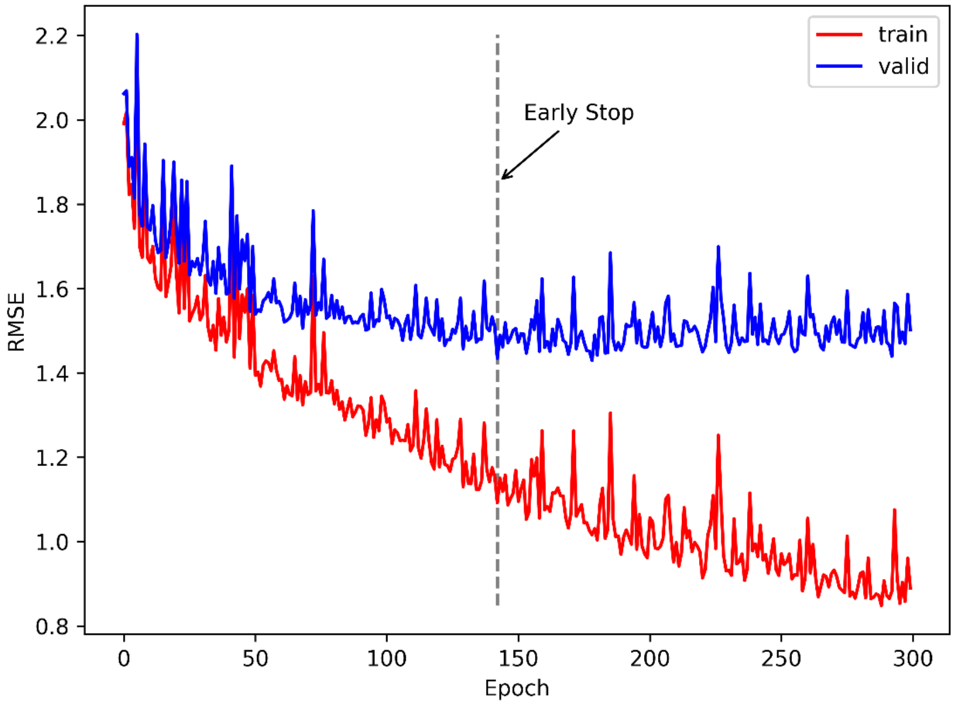

Figure S3. The RMSE loss of one-fold of the 5-fold cross-validation during hyperparameter search on protein–ligand binding affinity prediction task. Related to STAR Methods. Early stopping criterion is that the RMSE on validation set is no longer improving in 30 epochs. As shown in the figure, it will stop at about epoch 173 to avoid overfitting and get the best epoch 143.

Supplemental Tables

Table S1. Binding affinity prediction performance of two model variants of GraphDTA on the core set v.2016. Related to Table 1 and Figure 2A.

Model variants	Training dataset	RMSE	MAE	Pearson	SD	CI
GIN	Davis	5.375	4.925	0.053	2.174	0.501
GAT_GCN	Kiba	5.649	5.226	0.123	2.161	0.539

GraphDTA trained four different graph neural network variants on the Davis and Kiba dataset, respectively. And the GIN achieved the best performance on the Davis dataset while the GAT_GCN achieved the best performance on the Kiba dataset. So we examined both GIN and GAT_GCN on the core set v.2016, and chose the best scoring power (two key indicators: Pearson and SD) of these results as the performance of GraphDTA.

Table S2. Performance of two model variants of GraphDTA on the DUD-$\mathbf{E}_{\textbf{\textit{hand}}}$. Related to Figure 2B.

Model variant	Training dataset	Average AUC
GIN	Davis	0.468
GAT_GCN	Kiba	0.548

We examined both GIN and GAT_GCN on the DUD-E_{hand}, and chose the best average AUC of these results as the performance of GraphDTA. The reason for this is explained in Table S1.

Table S3. Performance of SEGSA_DTA with or without Hyperedge Convolution on the training set. Related to Figure 2C.

model	**Evaluation metrics** [mean ± 95% confidence interval]			
	$RMSE_{affinity}$	$Pearson_{affinity}$	$AUC_{interaction}$	$RMSE_{contribution}$
proEdge_DTA	1.397 [1.395-1.400]	0.698 [0.697-0.699]	0.722 [0.721-0.723]	0.476 [0.476-0.477]
noEdge_DTA	1.390 [1.386-1.393]	0.701 [0.700-0.702]	0.728 [0.726-0.730]	0.473 [0.472-0.474]
ligEdge_DTA	1.370 [1.365-1.372]	0.712 [0.711-0.713]	0.729 [0.728-0.730]	**0.462**[a] [0.461-0.463]
SEGSA_DTA	**1.343** [1.338-1.346]	**0.725** [0.724-0.727]	**0.744** [0.743-0.746]	**0.462** [0.461-0.463]

Bold indicates the best prediction performance.

[a] The p-values for all cases are less than 0.0001, except for the $RMSE_{contribution}$ of ligEdge_DTA (p-value 0.692 > 0.05).

Table S4. Performance of SEGSA_DTA with or without supervised attentions on the training set. Related to Figure 2D.

model	**Evaluation metrics** [mean ± 95% confidence interval]			
	$RMSE_{affinity}$	$Pearson_{affinity}$	$AUC_{interaction}$	$RMSE_{contribution}$
contriSA_DTA	1.396 [1.392-1.399]	0.699 [0.696-0.702]	0.491 [0.490-0.493]	0.488 [0.486-0.489]
noSA_DTA	1.387 [1.383-1.391]	0.705 [0.701-0.707]	0.473 [0.473-0.474]	0.515 [0.514-0.515]
interSA_DTA	1.386 [1.381-1.390]	0.704 [0.702-0.706]	0.716 [0.715-0.718]	0.512 [0.511-0.512]
SEGSA_DTA	**1.343** [1.338-1.346]	**0.725** [0.724-0.727]	**0.744** [0.743-0.746]	**0.462** [0.461-0.463]

Bold indicates the best prediction performance.

All p-values are less 0.0001.

Table S5. Summary of proteins used for the case study of mechanisms of selective binding of ligands to targets. Related to Figure 3.

Protein family	**Protein category**	**Ligand**	**Proteins for comparison**
COXs	Kinase	SC-558	COX-1, COX-2
5-HTs	G protein coupled receptor	CD10	5-HT1, 5-HT2
TREKs	ion channel	TKDC	TREK-1, TRAAK

Table S6. Summary of ligands used for the case study of guidance for structural-based lead optimization. Related to Figure 4.

Protein family	**Protein category**	**Protein**	**Ligands for comparison**
5-HTs	G protein coupled receptor	5-HT2	TKDC, 28NH
TREKs	ion channel	TRAAK	CD10, CD12

Table S7. SHAP values of ligands. Related to Figure 3 and Discussion.

Protein family	**Ligand**	**Protein**	**Inhibitory activity**	**Shap value of the ligand**
COXs	SC-558	COX-2_WT	**Strong**	**7.235**
		COX-2_V523I	**Weak**	**6.914**
5-HTs	CD10	5-HT2_WT	Weak	5.505
		5-HT2_M218T	Strong	5.350
TREKs	TKDC	TRAAK_WT	**Weak**	**3.608**
		TRAAK_E38T	**Medium**	**3.829**
		TRAAK_E38T_E41I	Strong	3.416

Table S8. Summary of ligand features. Related to STAR Methods.

Feature	Size	Description
Atom Feature		
atom symbol	9	[C, N, O, F, P, S, Cl, Br, I] (one-hot)
degree	4	[1, 2, 3, 4] (one-hot)
partial charge	1	Gasteiger Charges (float)
implicit hydrogen charge	1	the total charge for the implicit hydrogens (float)
hybridization	5	[sp, sp2, sp3, sp3d, other] (one-hot)
aromaticity	1	[0, 1] (one-hot)
hydrogens	4	[0, 1, 2, 3] (one-hot)
chirality	1	[0, 1] (one-hot)
Bond Feature		
bond type	4	[single, double, triple, aromatic] (one-hot)
conjugation	1	[0, 1] (one-hot)
ring	1	[0, 1] (one-hot)

Table S9. Summary of hyperparameter settings. Related to STAR Methods.

Parameter	optimal value	Description	Range of search
learning rate	3e-3	The learning rate	[1e-6, 1e-5, 1e-4, 3e-4, 1e-3, 3e-3, 0.01, 0.03, 0.1, 0.3, 1.0]
α	0.05	Loss weight of the non-covalent interaction prediction	[0.01, 0.03, 0.04, 0.045, 0.05, 0.055, 0.06, 0.07, 0.1, 0.2, 1, 10]
β	10.0	Loss weight of the residue contribution prediction	[2, 6, 7, 8, 9, 10, 11, 12, 13,14,15, 20]
layer node_fea	256	Layer nodes of feature extraction module	[128, 256, 512]
layer node	512	Layer nodes of prediction module	[64, 128, 256, 512, 1024]
dropout_fea	0.1	Dropout of feature extraction module	[0.1, 0.2, 0.3]
dropout	0.3	Dropout of prediction module	[0.1, 0.3, 0.5]
L2_weight_decoy	1e-4	The L2 regularization	[0, 1e-6, 1e-5, 1e-4, 1e-3, 1e-2, 1]

For the order of the hyperparameter search,

(1) The learning rate is first searched, as it is one of the most important hyperparameters in relation to the size and composition of the dataset and the parameter complexity of the model. All other hyperparameters are kept at a moderate value at this point.

(2) Next, the loss weight α and β are tuned.

(3) Then comes the network structure, using grid search to tune the number of layer nodes of both the feature extraction module and the prediction module. While the number of network layers is set empirically, the number of network layers for the feature extraction module is set to two layers with the same number of nodes (layer node_fea). The prediction module is a fully connected neural network set to contain two hidden layers, where the number of nodes in the second layer is half that of the first layer (layer node).

(4) Finally, the hyperparameters associated with the regularization term are tuned using a grid search to adjust the dropout of both the feature extraction module and the prediction module, followed by a search for the L2 regularization parameter.

Table S10. Summary of the PDBbind dataset preparation. Related to STAR Methods.

	Number of protein–ligand pairs	**Description of exclusion criteria**
Initial pairs	17, 679	--
Exclusion_1	11, 124	The data represented by Kd or Ki were selected.
Exclusion_2	8,728	The ligand requires the standard PDB ligand id and the corresponding binding affinity must be accurate and not a range value.
Exclusion_3	8,671	Discarded data where Ligand_ideal.pdb is empty or does not exist
Exclusion_4	7,261	The crystal structure resolution of the complex should be no greater than 2.5 Å
Exclusion_5	5,693	The ligand can be processed using RDKit, and its molecular weight must be less than 500.
Exclusion_6	5,629	In the calculation of non-covalent interactions, the complexes from the RCSB must contain the ligand corresponding to the binding activity record of the PDBbind; also, a total of 7 complexes that were identified as having no non-covalent bonding interactions were removed.
Exclusion_7	5,482	In the calculation of the contributions of residues, a total of 147 protein-ligand pairs were unable to be calculated.
Final pairs	5,482	--