Supporting Information

Predicting drug-target affinity by learning protein knowledge from biological networks

Wenjian Ma[1], *Shugang Zhang*[1,*], *Zhen Li*[2], *Mingjian Jiang*[3], *Shuang Wang*[4], *Nianfan Guo*[1]*,* *Yuanfei Li*[1], *Xiangpeng Bi*[1], *Huasen Jiang*[1], *Zhiqiang Wei*[1]

[1]College of Computer Science and Technology, Ocean University of China, Qingdao 266100, China

[2]College of Computer Science and Technology, Qingdao University, Qingdao University, Qingdao 266071, China

[3]School of Information and Control Engineering, Qingdao University of Technology, Qingdao 266033, China

[4]College of Computer Science and Technology, China University of Petroleum (East China), Qingdao 266580, China

*Corresponding author: Shugang Zhang (zsg@ouc.edu.cn)

This PDF file includes:

1 Supplementary Notes

Supplementary Note S1. The performance of the model with different number of GCN layers

The proposed method applied GCN in both protein representation learning and drug representation learning. Specifically, 3-layer GCN was used in the drug branch to extract the features of drug molecules, and 2-layer GCN was used in the VGAE framework of the protein branch to generate the latent representations of proteins. To explore the impact of varying the number of layers on the model performance, we considered them as hyperparameters and performed 5-fold cross validation experiments on the Davis dataset. Noted that the VGAE framework requires at least two GCN layers [1], with the first layer used for aggregating local neighboring features and the second used for generating the mean value and the standard deviation for the following decoding phase.

Figure S1 plots performances of models with different layer numbers in the drug branch. Generally, they achieved comparable performances in terms of both MSE (0.209, 0.208, and 0.208 for 1, 2, and 3 layers, respectively) and CI (0.893, 0.892, and 0.890). However, as the baseline models like GraphDTA and DGraphDTA all used 3-layer GCN in the drug branch, we adopted identical settings to them for a fair comparison.

Figure S2 demonstrates performances of models with different layer numbers in VGAE. According to the evaluation results, the model under two settings achieved very close performance in terms of CI, while the 2-layer GCN achieved better MSE score (0.208) compared with 3-layer GCN (0.215).

Supplementary Note S2. The performance of the model with protein-level features or residue-level features

To demonstrate the contributions of protein-level features and residue-level features to the model performance, we additionally performed the comparison experiments on Davis. Specifically, we used protein representations learned from residue-level sequence features and residue-level graph features, respectively, to replace those learned from protein-level features for DTA prediction. Among them, the residue-level sequence features only included 21 residue types (20 kinds of standard residue and 1 unknown residue), while the residue-level graph features contained not only residue types, but also residue-level biochemical features like residue weight and hydrophobicity (see Table S1) and the interaction between residues. On this basis, the protein representations learned from protein-level features, residue-level sequence features, and residue-level graph features were separately used for DTA prediction, the results of which are shown in Figure S3.

It can be seen from Figure S3 that the model with protein-level features achieved optimal performance compared with those with residue-level features, i.e., MSE = 0.196 and CI = 0.906. In addition, the model with residue-level sequence features obtained the worst performance, i.e., MSE = 0.257 and CI = 0.880.

Supplementary Note S3. The performance of model with different scales of biological networks

In this study, we constructed large SSN and PPI networks for all available human proteins from SwissProt and STRING, rather than only for the target proteins in a certain DTA dataset, to learn protein representations. On this basis, more protein nodes and edges that represent relations between the proteins can be included in the two large biological networks, and consequently more protein prior knowledge can be learned to generate expressive protein representations. Here, we implemented the comparison experiments on the Davis dataset to study the impact of the scale of biological networks. In detail, we built three groups of networks: (1) the networks with 18,552 protein nodes (i.e., networks-large); (2) the networks with only the target proteins in Davis as nodes (i.e., networks-targets), in which 31 'orphan nodes' were existed; (3) the networks without 'orphan nodes' (i.e., networks-targets (no orphan nodes)), which introduced other neighbor protein nodes outside the Davis to eliminate 'orphan nodes'. The model performances under these three settings are shown in Figure S4.

According to Figure S4, the model with 18,552 protein nodes achieved optimal performance compared with other two groups of networks, i.e., 0.194 vs. 0.197 vs. 0.200 (MSE) and 0.906 vs. 0.903 vs. 0.902 (CI).

Supplementary Table

Table S1. Residue-level biochemical features

	Residue-level features	Dimension
1	One-hot encoding of the residue symbol	21
2	Whether the residue is aliphatic	1
3	Whether the residue is aromatic	1
4	Whether the residue is polar neutral	1
5	Whether the residue is acidic charged	1
6	Whether the residue is basic charged	1
7	Residue weight	1
8	The negative of the logarithm of the dissociation constant for the –COOH group	1
9	The negative of the logarithm of the dissociation constant for the –NH3 group	1
10	The negative of the logarithm of the dissociation constant for any other group in the molecule	1
11	Position-specific scoring matrix (PSSM)	21
12	The pH at the isoelectric point	1
13	Hydrophobicity of residue (pH = 2)	1
14	Hydrophobicity of residue (pH = 7)	1

2 Supplementary Figures

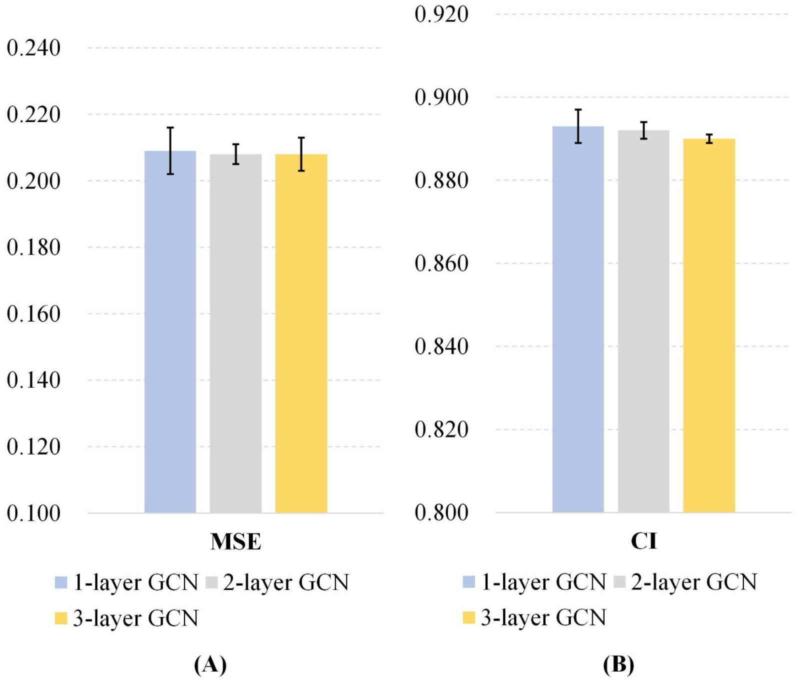


Figure S1. Performance of models with different GCN layers in the drug branch. (A) MSE (B) CI.

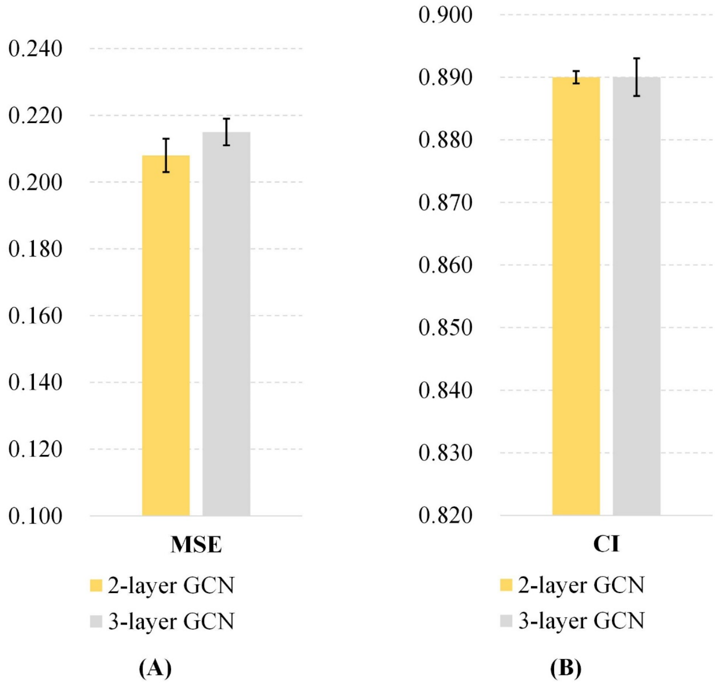


Figure S2. Performance of model with different layers in VGAE of the protein branch. (A) MSE (B) CI.

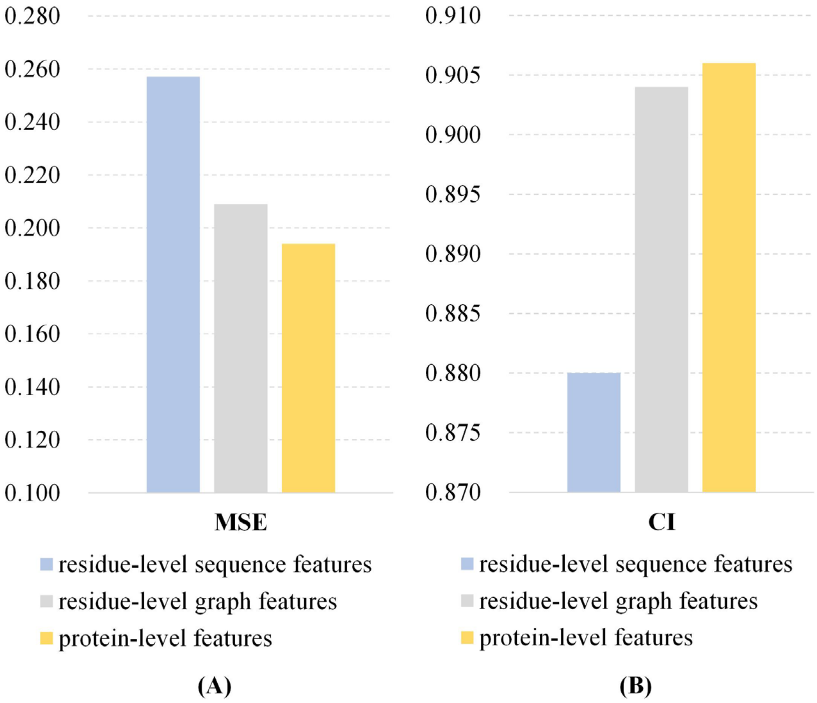
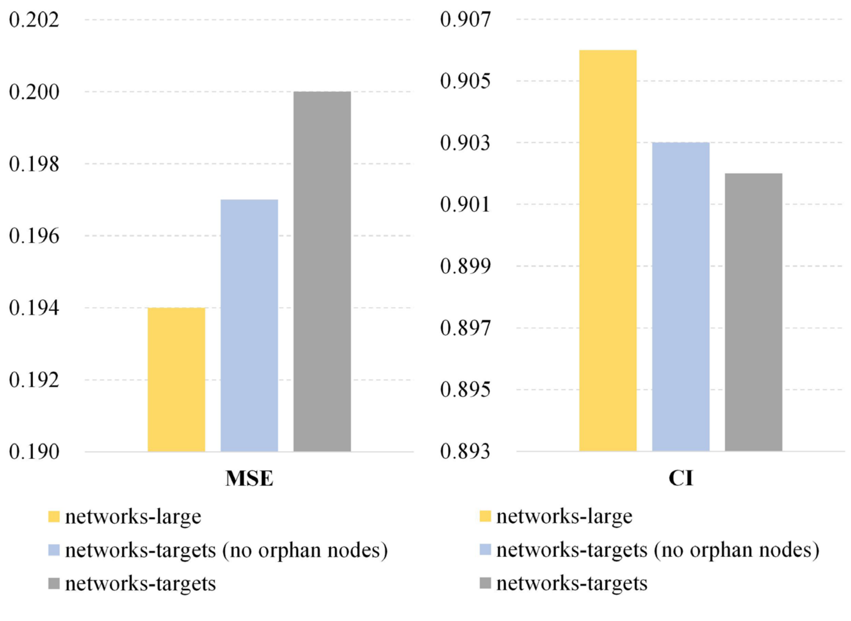


Figure S3. Performance of model with protein-level features or residue-level features in terms of (A) MSE and (B) CI



(A) (B)

Figure S4. Performance of model with three groups of networks in terms of (A) MSE (B) CI.

Supplementary References

[1] T. N. Kipf and M. Welling, “Variational graph auto-encoders,” *arXiv Prepr. arXiv1611.07308*, 2016.