HaPPy: Harnessing the Wisdom from Multi-Perspective Graphs for Protein-Ligand Binding Affinity Prediction (Student Abstract)

Xianfeng Zhang[1], Yanhui Gu[1]*, Guandong Xu[2], Yafei Li[3]*, Jinlan Wang[4], Zhenglu Yang[5]

[1]School of Computer and Electronic Information Science, Nanjing Normal University, Nanjing, China
[2]School of Computer Science and Advanced Analytics Institute, University of Technology Sydney, Sydeny, Australia
[3]School of Chemistry and Materials Science, Nanjing Normal University, Nanjing, China
[4]School of Physics, Southeast University, Nanjing, China
[5]College of Computer Science, Nankai University, Tianjin, China
{xf_zhang, gu}@njnu.edu.cn, guandong.xu@uts.edu.au, liyafei@njnu.edu.cn, jlwang@seu.edu.cn, yangzl@nankai.edu.cn

Abstract

Gathering information from multi-perspective graphs is an essential issue for many applications especially for protein-ligand binding affinity prediction. Most of traditional approaches obtained such information individually with low interpretability. In this paper, we harness the rich information from multi-perspective graphs with a general model, which abstractly represents protein-ligand complexes with better interpretability while achieving excellent predictive performance. In addition, we specially analyze the protein-ligand binding affinity problem, taking into account the heterogeneity of proteins and ligands. Experimental evaluations demonstrate the effectiveness of our data representation strategy on public datasets by fusing information from different perspectives.

Introduction

As an important step in the process of drug design, protein-ligand binding affinity prediction task has also attracted great attention in the fields of deep learning and biochemistry in recent years. By abstracting the spatial structure of protein-ligand complexes into different forms such as sequences, graphs, 3D grids and others, the existing prediction models have achieved good performance(Du et al. 2022). However, we found that these studies were only limited to certain parts of the complex structure, without comprehensive, multi-perspective and specialized analysis, which often led to poor interpretability, and is the last thing to face in the process of drug development.

The core of this task is to capture different influence factors of binding affinity from the structure of complex, such as intramolecular and intermolecular forces of ligand and protein, local information of binding pocket and complete protein structure information. Therefore, we propose HaPPy strategy by **Ha**rnessing the wisdom from multi-**P**erspective graphs for **P**rotein-ligand binding affini**y** prediction. Specifically, we represent the spatial structure of the complex into four parts: ligand graph, protein pocket graph, interactive bipartite graph, and protein sequence, and then use different encoders to encode them, attempting to extract different features that affect binding affinity, the overall framework of our strategy is shown in Figure 1. Through the comprehensive utilization of information from different perspectives, our model provides a comprehensive description of protein-ligand complex structures, which then demonstrates competitive predictive performance and reflects the effects of different information.

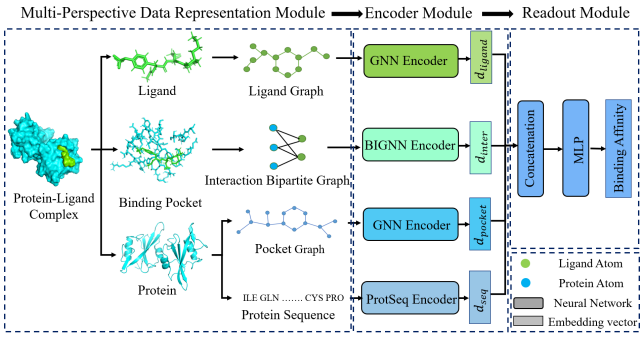

Figure 1: The overall framework of our strategy

Strategy

The key of our strategy is to find some data representation structures for each part of the protein-ligand complex that accurately describes the properties, then harness the advantage from different perspectives.

Homogeneous Graph Neural Network Encoder It is a very intuitive and major approach to encode a molecule as a graph, atoms as nodes, and chemical bonds as edges. We represent the ligand and the partial protein structure in the binding pocket as two undirected graphs G_{ligand} and G_{pocket}. The reason why the complete protein structure can not be represented as a graph is that it contains thousands of atoms and chemical bonds, which will result in a huge resource consumption. And the protein structure in the binding pocket has the most important impact on ligand binding. For the ligand graph and the protein pocket graph, the nodes in the graph follow the same distribution, so the current major

*Corresponding author: Yanhui Gu, Yafei Li.

homogeneous graph neural network(Kipf and Welling 2016) can be used to encode the entire graph. Simply put, this type of homogeneous graph neural network iteratively updates its own node information by aggregating the feature information of neighbor nodes and edges. For the node u, the form can be described as:

$$h_u^k = U(h_u^{k-1}, AG(M(h_v^{k-1}, h_u^{k-1}, e_{uv}))), v \in \mathcal{N}(u) \quad (1)$$

where h_u^k is the feature vector of the k-th layer, and $\mathcal{N}$ is the set of neighbors of u. And it contains three functions, update function (U), aggregation function(AG), and message passing function(M). By passing information between atoms within a molecule, we can naturally capture their intramolecular forces.

Protein Sequence Encoder In order to reduce the limitation of considering only part of the protein structure, we use pre-trained protein language model (Elnaggar et al. 2020) to obtain global information of the protein from amino acid sequence, because protein is essentially an amino acid sequence, and different amino acid sequences determine different protein structures.

Bipartite Graph Neural Network Encoder We construct an interactive bipartite graph G_{inter} based on the cutoff distance between atoms to capture the interactive information. It is precisely because bipartite graphs discard information about the relationships between nodes belonging to the same set, so that it only focuses on the interactive information between different sets. For bipartite graphs, the nodes in different sets follow different distributions and contain different properties. Therefore, message passing needs to be performed on two groups of nodes respectively, which can be described as:

$$h_u^k = U_u(h_u^{k-1}, AG(M_u(h_v^{k-1}, h_u^{k-1}, e_{uv}))), v \in \mathcal{N}(u) \quad (2)$$

$$h_v^k = U_v(h_v^{k-1}, AG(M_v(h_u^{k-1}, h_v^{k-1}, e_{vu}))), u \in \mathcal{N}(v) \quad (3)$$

In addition, to obtain the global embedding of bipartite graphs focusing on intermolecular forces, we do pooling at the edge level rather than pooling directly at the node level.

Readout Module Through different encoders, we obtain four embedding vectors from different perspectives, taking into account different information. These four embedding vectors are concatenated to obtain the embedding vector of the entire protein-ligand complex, which is sent to the multilayer perceptron to predict the affinity.

Experimental Evaluation

We use the same dataset and training strategy as the baseline models (Wang et al. 2021), and the comparison results are shown in table 1. It can be found that the prediction results of our model reach a very good level, and our MAE value is reduced by 3.7% relative to the best baseline model.

In addition, through the experimental results in Figure 2, it can be found that after removing the information from some perspectives, the prediction performance of the model decreases, and the information of different perspectives has its own role. In conclusion, our model achieves the best performance when constructing graphs at the atomic level and considering information from all perspectives.

Model	MAE↓	RMSE↓	SD↓	R↑
Pafncy	1.129	1.418	1.375	0.775
DeepDTA	1.148	1.443	1.445	0.749
DeepDTAF	1.073	1.355	1.337	0.789
OnionNet	0.983	1.287	1.282	0.781
DPLA	0.972	1.255	1.248	0.820
Ours	**0.936**	**1.228**	**1.221**	**0.827**

Table 1: Result on PDBbind_v2016 core set

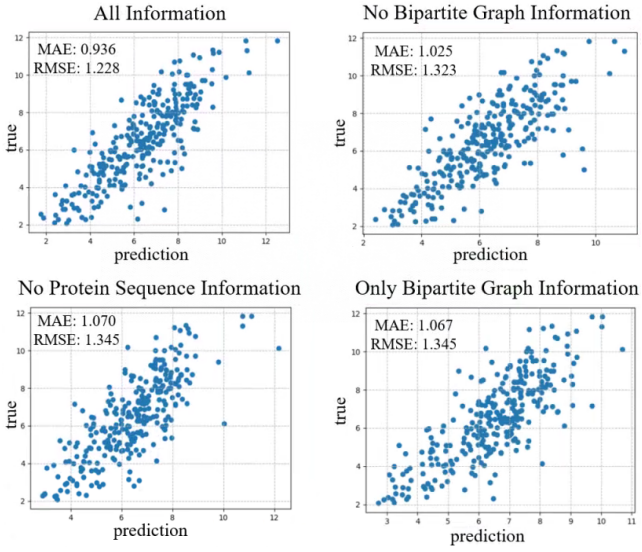

Figure 2: The influence from different perspectives

Acknowledgments

This work is supported by National Natural Science Foundation of China (Grant No.22033002 and No.21873050).

References

Du, B.-X.; Qin, Y.; Jiang, Y.-F.; Xu, Y.; Yiu, S.-M.; Yu, H.; and Shi, J.-Y. 2022. Compound–protein interaction prediction by deep learning: Databases, descriptors and models. *Drug Discovery Today*, 27(5): 1350–1366.

Elnaggar, A.; Heinzinger, M.; Dallago, C.; Rihawi, G.; Wang, Y.; Jones, L.; Gibbs, T.; Feher, T.; Angerer, C.; Steinegger, M.; et al. 2020. ProtTrans: towards cracking the language of Life's code through self-supervised deep learning and high performance computing. *arXiv preprint arXiv:2007.06225*.

Kipf, T. N.; and Welling, M. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.

Wang, W.; Sun, B.; Liu, D.; Wang, X.; and Zhang, H. 2021. DPLA: prediction of protein-ligand binding affinity by integrating multi-level information. In *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 3428–3434.