Improving Drug-Target Interaction Predictions Through an Explainable Graph Transformer Model

Baozhong Zhu [1] and Hongjie Wu [1]

[1] Suzhou University of Science and Technology
hongjie.wu@qq.com

Abstract. Drug discovery is a complex and time-consuming process. Identifying drug-target interactions (DTIs) is crucial for early-stage drug development. This study introduces a novel model for DTI prediction that leverages protein binding sites and self-attention mechanisms. The model achieves high performance in DTI prediction and provides interpretability by identifying protein regions interacting with ligands.

Keywords: Drug-target Interaction Prediction.

1 Introduction

Drug discovery involves identifying drug-target interactions, which is a complex and resource-intensive process. Computational methods have been proposed to facilitate DTI identification and expedite drug discovery. This study presents a novel architecture for DTI prediction using protein binding sites and self-attention mechanisms.

Drug discovery is a complex and time-consuming process, and despite significant investments, success rates remain suboptimal [1]. Proteins are the primary targets of drugs and the identification of drug-target interaction (DTI) has become a crucial task in early-stage drug development and drug repurposing [2]. Since experimental DTI studies are expensive and time-consuming, computational methodologies have been proposed to facilitate the identification of putative DTI, thereby expediting the process of drug discovery [3]. One of the main methods for virtual screening involves predicting potential drugs by screening out drug candidate ligands for receptor proteins of interest from large-scale compound ligand libraries using many calculations [4]. Virtual screening methods can be divided into two categories: receptor-based virtual screening and ligand-based virtual screening. Receptor-based virtual screening mainly studies the three-dimensional structure of proteins and seeks interactions with small molecule compounds from the three-dimensional structure, making it also known as structure-based virtual screening [5]. However, these methods have practical limitations due to their heavy reliance on the high-quality three-dimensional structure of proteins and their computational expenses and inefficiencies. Ligand-based virtual screening typically begins with ligands and analyzes molecular structure and activity information of known inhibitors to summarize structural features that significantly contribute to their binding

capacity. This learned knowledge is then used to screen new ligands to find compound molecules that meet the requirements. Virtual screening methods often rely on predicting drug-target interactions, which can be understood as a series of continuous values that express the intensity of different drug-target interactions.

With the rapid development of deep learning methods [6], researchers have used deep learning models to measure drug-target interactions as binary classification tasks[7]. These DTI prediction models have been hugely successful because they can automatically capture data depth features, resulting in better models with excellent capabilities in complex molecular data processing [8]. DTI's deep learning models can be divided into two main categories [9]. One type act on processing sequence-based representations of input data, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs). In related works, Peng proposed a method based on convolutional neural networks to extract drug and protein features from heterogeneous networks, and used convolutional neural network models to predict the interaction between drugs and proteins [10]. Karimi proposes a semi-supervised deep learning model that unifies recurrent and convolutional neural networks to jointly encode molecular representation and predict affinity using unlabeled and labeled data [11].

However, these models usually express drugs in the form of strings, and one-dimensional sequences are not a natural way of expressing molecules. Therefore, to compensate for the lack of molecular structure information, a second type of deep learning model, the graph neural network (GNN), was introduced, and the use of graph convolutional networks has also proven to be more beneficial for computational drug discovery [12]. GNN uses a graphical description of molecules, where atoms and chemical bonds correspond to nodes and edges, respectively [13]. The most commonly used GNN-based models today are the graph convolutional neural network (GCNN) [14] and the graph attention network (GAT), which is one of the variants of GCNN. Related work includes Zhao using the constructed graph convolutional network to learn the drug-protein pairs built to improve the prediction accuracy [15]. Zhao proposes a new graph convolutional DTI prediction model. Specifically, the first-order neighbor information of a node can be aggregated through GCN; The high-order neighbor information of the node is learned by the graph embedding method, which improves the accuracy of prediction [16].

Despite the impressive performance of both CNN-based and graph-based neural network methods in DTI prediction, certain challenges remain unresolved [17]. One significant limitation of most deep learning methods is that they employ only a few CNN layers, resulting in the compression of all feature information into a small area, which may cause the loss of local features of the original data. Moreover, all graph-based models are currently represented using the amino acid sequence of the protein, which cannot capture the crucial 3D structural features that are essential in DTI prediction.

Obtaining a high-resolution 3D structure of a protein is a difficult task due to its complex nature and large number of atoms, necessitating a massive 3D (sparse) matrix to capture the entire structure. This paper proposes a novel approach for predicting DTI that leverages the structural features of small molecules and protein binding sites in the form of graphs. To preserve the influence of molecular structure on the prediction results, a transformer model is introduced to extract global features. Moreover, a

selfattention Bidirectional Long Short-Term Memory mechanism is employed to identify the parts of the protein that are most likely to bind to a given drug, thereby enhancing the model's interpretability

2 Method

The proposed framework consists of four main modules. The Data Preparation module extracts protein binding sites. The Graph Embedding Learning module generates a graph map of protein pockets and ligands using TAGCN to extract global and local features. The Feature Extraction module uses a transformer block and a Self-attentive BiLSTM block to learn the relationship between ligands and protein binding sites. The Prediction module uses a binary classifier for DTI prediction.

2.1 Framework

The framework includes a pretreatment module for identifying protein binding sites, a graph representation module, a feature extraction module with a transformer block and BiLSTM block, and a prediction module for DTI prediction.

TAGCN is used to generate embeddings from the graph representation, capturing both local and global features.The transformer block focuses on global information, while the Self-attentive BiLSTM block identifies key contributors to predicted interactions. A two-layer fully connected neural network with a logistic sigmoid function predicts DTI probabilities.

Our model was subjected to rigorous evaluation using two widely recognized DTI datasets, namely, the DUD-E dataset and the Human dataset. These benchmark datasets are commonly used in the field of drug target interaction prediction. The DUD-E dataset comprises 102 targets belonging to eight distinct protein families. Each target comprises roughly 224 active compounds and more than 10,000 bait molecules. On the other hand, the Human dataset was constructed by combining a highly credible and reliable set of negative drug-protein samples with known positive samples using systematic in silico screening methods. The dataset contains 5423 interactions between drug and target molecules. Table 1 presents a summary of the key statistics for these two datasets. All datasets are publicly available. DUD-E dataset is available at http://dude.docking.org, Human dataset is available at https://github.com/IBMInterpretableDTIP.

2.2 Train

The model was evaluated using the DUD-E dataset and the Human dataset, which are standard benchmarks for DTI prediction.

Proteins are represented as graphs with atoms as nodes and connections as edges. Ligands are represented in SMILE format and encoded similarly.

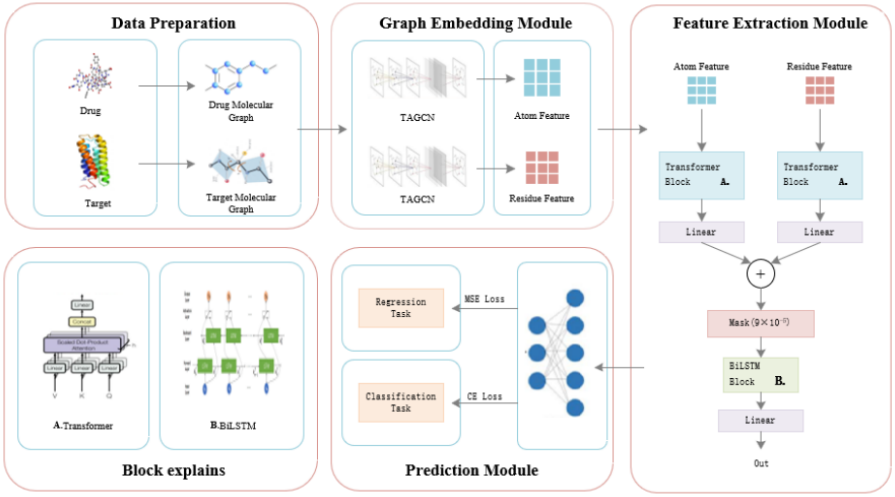

Fig. 1. Architecture.

3 Results

The model demonstrated superior performance over existing DTI predictive models. It achieved high AUC scores, precision, recall, and F1 scores on the Human dataset and showed significant enrichment in the DUD-E dataset comparison. The model's improved performance can be attributed to sophisticated input representations, robust feature extraction mechanisms, and the interpretability provided by the self-attention mechanism. The proposed model effectively captures drug-target interactions and provides interpretability by identifying specific protein binding sites. This approach holds promise for accelerating drug discovery processes.

We posit that the improved performance of our proposed model can be attributed to several factors:

(1) Input representation plays a crucial role in predicting the binding affinity of drug-target complexes. Utilizing more sophisticated input representations, such as structural diagrams, can aid in capturing crucial structural information regarding molecules.

(2) Feature extraction technique is an important consideration, and transformerbased architectures provide a robust automatic feature extraction mechanism that can capture high-order nonlinear relationships. Additionally, graph-based neural networks that employ graphical representations of drugs and proteins can effectively capture the topological relationships between drug molecules and target proteins, further enhancing the performance.

(3) To more effectively model and interpret the binding relationship of drug-target complexes, we introduce a self-attentive BiLSTM with masks. This model not only retains past and future information of the sequence input flowing in both directions but also explicates the degree of binding of drug-target complexes through the attention weight ratio.

Acknowledgments. This research was supported by the National Natural Science Foundation of China and other institutions.

Disclosure of Interests. The authors declare no conflict of interest.

References

1. Abbasi Mesrabadi, H., Faez, K. & Pirgazi, J. Drug–target interaction prediction based on protein features, using wrapper feature selection. Sci Rep 13, 3594 (2023). https://doi.org/10.1038/s41598-023-30026-y.
2. Soh, J., Park, S. & Lee, H. HIDTI: integration of heterogeneous information to predict drug target interactions. Sci Rep 12, 3793 (2022). https://doi.org/10.1038/s41598-022-07608-3.
3. Azuaje, F., Zhang, L., Devaux, Y. et al. Drug-target network in myocardial infarction reveals multiple side effects of unrelated drugs. Sci Rep 1, 52 (2011). https://doi.org/10.1038/srep00052.10
4. Beroza, P., Crawford, J.J., Ganichkin, O. et al. Chemical space docking enables large-scale structure-based virtual screening to discover ROCK1 kinase inhibitors. Nat Commun 13, 6447 (2022). https://doi.org/10.1038/s41467-022-33981-8.
5. Crunkhorn, S. Novel virtual screening approach. Nat Rev Drug Discov 16, 18 (2017). https://doi.org/10.1038/nrd.2016.272.
6. Ding, Yijie, Jijun Tang, and Fei Guo. "Identification of drug–target interactions via dual laplacian regularized least squares with multiple kernel fusion." Knowledge-Based Systems 204 (2020): 106254.
7. Ding, Yijie, Jijun Tang, and Fei Guo. "Identification of drug–target interactions via fuzzy bipartite local model." Neural Computing and Applications 32 (2020): 10303-10319.
8. Ding, Yijie, Jijun Tang, and Fei Guo. "Identification of drug-side effect association via semisupervised model and multiple kernel learning." IEEE journal of biomedical and health informatics 23.6 (2018): 2619-2632.
9. Ding, Yijie, Jijun Tang, and Fei Guo. "Identification of drug-target interactions via multiview graph regularized link propagation model." Neurocomputing 461 (2021): 618-631.
10. Peng J, Li J, Shang X. A learning-based method for drug-target interaction prediction based on feature representation learning and deep neural network. BMC Bioinformatics. 2020 Sep 17;21(Suppl 13):394. doi: 10.1186/s12859-020-03677-1. PMID: 32938374; PMCID: PMC7495825.
11. Karimi M, Wu D, Wang Z, Shen Y. DeepAffinity: interpretable deep learning of compound protein affinity through unified recurrent and convolutional neural networks. Bioinformatics. 2019 Sep 15;35(18):3329-3338. doi: 10.1093/bioinformatics/btz111. PMID: 30768156; PMCID: PMC6748780.
12. Veličković, Petar, et al. "Graph attention networks." arXiv preprint arXiv:1710.10903 (2017).
13. Torng, Wen, and Russ B. Altman. "Graph convolutional neural networks for predicting drug-target interactions." Journal of chemical information and modeling 59.10 (2019): 4131- 4149.

14. Lim, Jaechang, et al. "Predicting drug–target interaction using a novel graph neural network with 3D structure-embedded graph representation." Journal of chemical information and modeling 59.9 (2019): 3981-3988.
15. Tianyi Zhao, Yang Hu, Linda R Valsdottir, Tianyi Zang, Jiajie Peng, Identifying drug–target interactions based on graph convolutional network and deep neural network, Briefings in Bioinformatics, Volume 22, Issue 2, March 2021, Pages 2141–2150, https://doi.org/10.1093/bib/bbaa044
16. Zhao BW, You ZH, Hu L, Guo ZH, Wang L, Chen ZH, Wong L. A Novel Method to Predict Drug-Target Interactions Based on Large-Scale Graph Representation Learning. Cancers (Basel). 2021 Apr 27;13(9):2111. doi: 10.3390/cancers13092111. PMID: 33925568; PMCID: PMC8123765.
17. Ding, Yijie, Jijun Tang, and Fei Guo. "Protein crystallization identification via fuzzy model on linear neighborhood representation." IEEE/ACM transactions on computational biology and bioinformatics 18.5 (2019): 1986-1995.
18. Saberi Fathi, Seyed Majid, and Jack A. Tuszynski. "A simple method for finding a protein's ligand-binding pockets." BMC Structural Biology 14.1 (2014): 1-9.
19. Ding, Yijie, Jijun Tang, and Fei Guo. "Human protein subcellular localization identification via fuzzy model on kernelized neighborhood representation." Applied Soft Computing 96 (2020): 106596.11
20. Wu, Hongjie, et al. "Empirical potential energy function toward ab initio folding G protein coupled receptors." IEEE/ACM Transactions on Computational Biology and Bioinformatics 18.5 (2020): 1752-1762.
21. Wang, Hao, et al. "Exploring associations of non-coding RNAs in human diseases via three matrix factorization with hypergraph-regular terms on center kernel alignment." Briefings in Bioinformatics 22.5 (2021): bbaa409.
22. Yang, Hongpeng, et al. "Drug–disease associations prediction via multiple kernel-based dual graph regularized least squares." Applied Soft Computing 112 (2021): 107811.
23. Sun, Mengwei, et al. "MLapSVM-LBS: Predicting DNA-binding proteins via a multiple Laplacian regularized support vector machine with local behavior similarity." Knowledge Based Systems 250 (2022): 109174.
24. Du, Jian, et al. "Topology adaptive graph convolutional networks." arXiv preprint arXiv:1710.10370 (2017).
25. Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems 30 (2017).
26. Zhou, Peng, et al. "Attention-based bidirectional long short-term memory networks for relation classification." Proceedings of the 54th annual meeting of the association for computational linguistics (volume 2: Short papers). 2016.
27. Yazdani-Jahromi, Mehdi, et al. "AttentionSiteDTI: an interpretable graph-based model for drug-target interaction prediction using NLP sentence-level relation classification." Briefings in Bioinformatics 23.4 (2022): bbac272.
28. Liu, Hui, et al. "Improving compound–protein interaction prediction by building up highly credible negative samples." Bioinformatics 31.12 (2015): i221-i229.
29. Kipf, Thomas N., and Max Welling. "Semi-supervised classification with graph convolutional networks." arXiv preprint arXiv:1609.02907 (2016).

30. Wang, Erniu, et al. "A graph convolutional network–based method for chemical-protein interaction extraction: algorithm development." JMIR Medical Informatics 8.5 (2020): e17643.
31. Wu, Yifan, et al. "BridgeDPI: a novel graph neural network for predicting drug–protein interactions." Bioinformatics 38.9 (2022): 2571-2578.
32. Durrant, Jacob D., and J. Andrew McCammon. "NNScore 2.0: a neural-network receptor– ligand scoring function." Journal of chemical information and modeling 51.11 (2011): 2897- 2903.
33. Ballester, Pedro J., and John BO Mitchell. "A machine learning approach to predicting protein–ligand binding affinity with applications to molecular docking." Bioinformatics 26.9 (2010): 1169-1175.
34. Ragoza, Matthew, et al. "Protein–ligand scoring with convolutional neural networks." Journal of chemical information and modeling 57.4 (2017): 942-957.
35. Torng, Wen, and Russ B. Altman. "Graph convolutional neural networks for predicting drug-target interactions." Journal of chemical information and modeling 59.10 (2019): 4131- 4149.
36. Tsubaki, Masashi, Kentaro Tomii, and Jun Sese. "Compound–protein interaction prediction with end-to-end learning of neural networks for graphs and sequences." Bioinformatics 35.2 (2019): 309-318.
37. Wang, Erniu, et al. "A graph convolutional network–based method for chemical-protein interaction extraction: algorithm development." JMIR Medical Informatics 8.5 (2020): e17643.