Leveraging Local Protein Structures for Enhanced Drug-Target Binding Affinity Predictions Using Deep Learning Techniques

Runhua Zhang [1] and Hongjie Wu [1]

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

Abstract. The traditional drug discovery process is both time-consuming and costly. Utilizing artificial intelligence to predict drug-target binding affinity (DTA) has become a crucial approach for accelerating new drug discovery. This study introduces a novel deep learning-based method that incorporates both the primary and secondary structures of proteins to better represent the local and global features of proteins. We employ convolutional neural networks (CNNs) and graph neural networks (GNNs) to model proteins and drugs separately, capturing their interactions more effectively. Our method demonstrated improved performance in predicting DTA compared to state-of-the-art methods on two benchmark datasets.

Keywords: Drug-Target Binding Affinity Prediction.

1 Introduction

Developing a new drug that reaches the market costs approximately $2.6 billion, with a low approval rate of less than 12%. Therefore, computer-aided drug development has become a hot research topic. Accurately identifying drug-target interactions is essential in the computational stages of drug development. Our method focuses on predicting DTA by representing proteins using both their primary and secondary structures.

Developing a new drug that can be brought to market costs approximately $2.6 billion, and the approval rate of new drugs that enter clinical trials is less than 12% [1,2]. Moreover, developing a new drug requires a significant amount of time [3]. Therefore, computer-aided drug development has become a hot research topic in recent years [4]. Accurately identifying drug-target interactions is an essential step in the computational stages of drug development [5]. Currently, there are mainly two categories of computational methods used for predicting drug-target interactions. The first type treats interaction prediction as a binary classification task [6], that is, determining whether a drug and a target interact or not. The other type treats it as a regression task for predicting the binding affinity between the drug and the target. Binding affinity can measure the strength of drug-target interactions, and is usually expressed using inhibition constant (Ki), dissociation constant (Kd), or the half maximal inhibitory concentration (IC50) [7].

Our method focuses mainly on predicting drug-target binding affinity (DTA).

There are several computational methods used for predicting DTA. One approach is the ligand-based method which compares a query ligand to known ligands based on its target protein. However, if the number of known ligands for the target protein is insufficient [8], the predictions may be unreliable [9]. Another approach is molecular docking [10], which models the binding of compounds and proteins in conformational space based on their 3D structures. However, preparing 3D protein-ligand complexes can be quite challenging [11].

Predicting DTA using computational methods typically involves three main steps.

First, drug and target protein data are converted into computationally ready vectors or graphs using various encoding methods [12]. The commonly used representation forms of drugs mainly include simplified molecular linear input specification (SMILES) [13], molecular fingerprint and graph. Proteins are usually represented using one-hot encoding to capture their primary sequences. Second, different feature extraction methods are applied to obtain representative features of drugs and proteins, which are then used to replace their original input features. Finally, a regression process is performed to combine the respective representations and predict binding affinities.

In recent years, deep learning (DL) has made significant progress in the field of computer-aided drug design [14], particularly in the prediction of DTA. Many DLbased methods have been developed to improve DTA prediction performance. One of the earliest DL-based DTA prediction models, DeepDTA [15] uses one-dimensional (1D) convolutional neural networks (CNN) to extract sequence features of drugs and proteins, it uses the protein primary sequence and the SMILES string of the drug ligand as input, without incorporating any additional input information. WideDTA [16] improves prediction performance by incorporating protein domain information. However, expressing drugs as SMILES strings leads to the loss of their original graph structure, motivating the use of graph neural networks (GNN). GraphDTA [17] represents drugs as graphs, using multiple GNN variants such as the graph convolutional network (GCN)[18], the graph attention network (GAT) [19], and the graph isomorphism network (GIN) [20], and retaining CNN to represent proteins. This model outperformed existing 1D methods, highlighting the importance of structural information. However, these models only consider the overall interaction between drugs and proteins. MGraphDTA[21] introduces dense connections into the GNN and builds an ultra-deep network structure consisting of 27 layers of GCN. This architecture enables the simultaneous capture of local and global structures of compounds, improving the prediction performance of DTA. Additionally, MGraphDTA proposes a new visualization method to better understand the role of GNN in DTA prediction. DeepAffinity [22] introduces an attention mechanism to learn the binding site information between compounds and proteins, improving model interpretability. These approaches have demonstrated the success of using CNNs for feature extraction from protein sequences. GraphDTA, on the other hand, uses a graph structure to represent drugs and applies GCN for feature extraction, leading to improved prediction performance. This indicates that graph structures can be effectively utilized in DTA prediction. The above method mainly uses the primary structure of the protein, that is, the amino acid sequence to represent and input,

and can only extract the global features of the protein, ignoring the local features of the protein in a segment.

In this paper, we propose a novel deep learning-based method for predicting DTA that integrates both global and local features of proteins. The entire model comprises three distinct modules: the global protein features module, the local protein featuresmodule, and the ligand module. The protein data is one-dimensional and consists of the amino acid sequence structure and secondary structure of the protein, while the drug ligand is represented using graph data. We use CNN to learn the representation of protein primary and secondary sequences, employ GAT and GCN to learn the graph data representation of drugs, and finally concatenate the features obtained from the convolution and maximal pooling layers of the three modules and fed them into the classification component.

2 Method

We evaluated our model on two public datasets: Davis and KIBA. Protein secondary structure information was predicted using MLRC methods and incorporated into our model. Primary protein sequences were represented using one-hot encoding, and secondary structures were represented using 8D one-hot vectors. Drugs were represented as graphs using SMILES strings converted with RDKit.

Our model predicts drug-target interactions as a regression task, aiming to predict specific binding affinities. The proposed model architecture consists of three functional modules: global protein features, local protein features, and ligand features.

2.1 Global Protein Features Module

This module uses a CNN to learn the representation of protein primary sequences. The primary sequence is represented as a one-dimensional sequence of amino acids using a 20D one-hot encoding scheme. The CNN consists of three convolutional layers with an increasing number of filters, followed by max pooling layers. This allows the model to capture global features of the protein sequence.

Local Protein Features Module. This module also uses a CNN, but it focuses on learning the representation of protein secondary structures. Secondary structures are represented using an 8D one-hot vector for each amino acid type. The CNN here also consists of three convolutional layers followed by max pooling layers, capturing the local features of the protein.

Ligand Features Module. Drug compounds are represented as graphs with nodes representing atoms and edges representing chemical bonds. We use the Graph Attention Network (GAT) and Graph Convolutional Network (GCN) to learn the graph data representation of drugs. The GAT layer learns the importance of each node using a self-attention mechanism, while the GCN layers capture the connectivity relationship between graph nodes.

Classification Component. The features from the max pooling layers of the three modules are concatenated and fed into a classification component. This component

consists of three fully connected layers with 1024, 512, and 512 nodes, respectively. Dropout layers with a rate of 0.2 are used after the first two fully connected layers to prevent overfitting. The output layer predicts the binding affinity.

2.2 Train

The model was trained for 1000 epochs with a batch size of 512 and a learning rate of 0.0005. The Adam optimization algorithm and Rectified Linear Unit (ReLU) activation function were used to train the network. Mean Squared Error (MSE) was used as the loss function.

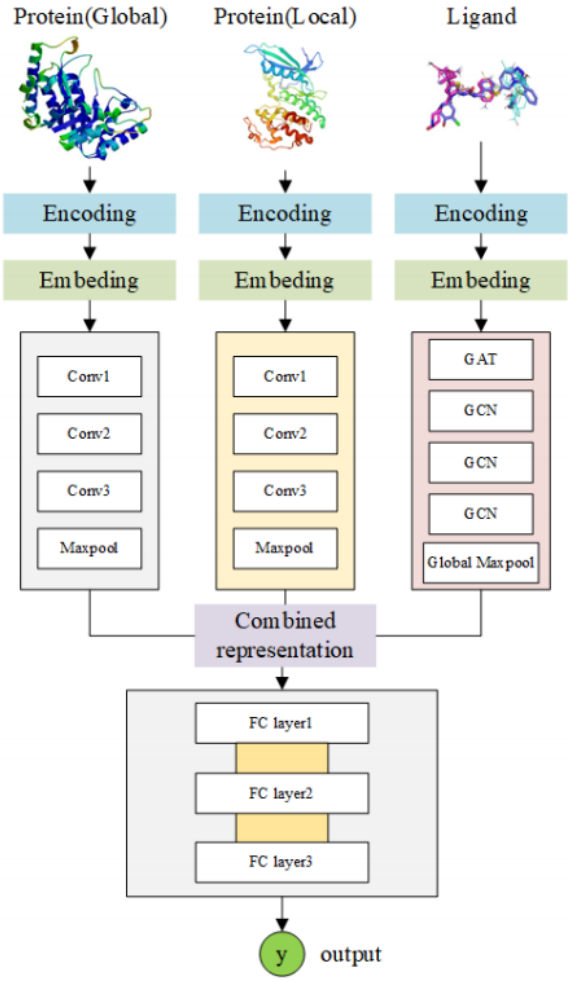

Fig. 1. Architecture.

3 Results

Our model showed superior performance compared to DeepDTA, WideDTA, GraphDTA, and AttentionDTA models on both the Davis and KIBA datasets. The inclusion of protein local features improved prediction accuracy, as indicated by the lower MSE and higher CI scores.

Our deep learning model, which incorporates primary and secondary protein structures, predicts drug-target binding affinity more accurately. The enriched datasets can be used in future experiments.

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. DiMasi J A, Grabowski H G, Hansen R W. Innovation in the pharmaceutical industry: new estimates of R&D costs[J]. Journal of health economics, 2016, 47: 20-33.
2. Mullard A. New drugs cost US $2.6 billion to develop[J]. Nature reviews. Drug discovery, 2014, 13(12): 877.
3. Ding Y, Tang J, Guo F. Identification of drug–target interactions via dual laplacian regular
ized least squares with multiple kernel fusion[J]. Knowledge-Based Systems, 2020, 204: 106254.
4. Sun M, Tiwari P, Qian Y, et al. MLapSVM-LBS: Predicting DNA-binding proteins via a multiple Laplacian regularized support vector machine with local behavior similarity[J]. Knowledge-Based Systems, 2022, 250: 109174.9
5. Ding Y, Tang J, Guo F. Identification of drug–target interactions via fuzzy bipartite local model[J]. Neural Computing and Applications, 2020, 32: 10303-10319.
6. Yamanishi Y, Kotera M, Kanehisa M, et al. Drug-target interaction prediction from chemical, genomic and pharmacological data in an integrated framework[J]. Bioinformatics, 2010, 26(12): i246-i254.
7. Tang J, Szwajda A, Shakyawar S, et al. Making sense of large-scale kinase inhibitor bioactivity data sets: a comparative and integrative analysis[J]. Journal of Chemical Information and Modeling, 2014, 54(3): 735-743.
8. Yang H, Ding Y, Tang J, et al. Drug–disease associations prediction via multiple kernelbased dual graph regularized least squares[J]. Applied Soft Computing, 2021, 112: 107811.
9. Ding Y, Tang J, Guo F. Human protein subcellular localization identification via fuzzy model on kernelized neighborhood representation[J]. Applied Soft Computing, 2020, 96: 106596.
10. Wu H, Ling H, Gao L, et al. Empirical potential energy function toward ab initio folding G protein-coupled receptors[J]. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2020, 18(5): 1752-1762.
11. Karimi M, Wu D, Wang Z, et al. Explainable deep relational networks for predicting compound–protein affinities and contacts[J]. Journal of chemical information and modeling, 2020, 61(1): 46-66.

12. Ding Y, Tang J, Guo F. Identification of drug-target interactions via multi-view graph regularized link propagation model[J]. Neurocomputing, 2021, 461: 618-631.
13. Weininger D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules[J]. Journal of chemical information and computer sciences, 1988, 28(1): 31-36.
14. Ding Y, Tang J, Guo F. Identification of drug-side effect association via semisupervised model and multiple kernel learning[J]. IEEE journal of biomedical and health informatics, 2018, 23(6): 2619-2632.
15. Öztürk H, Özgür A, Ozkirimli E. DeepDTA: deep drug–target binding affinity prediction[J]. Bioinformatics, 2018, 34(17): i821-i829.
16. Öztürk H, Ozkirimli E, Özgür A. WideDTA: prediction of drug-target binding affinity[J]. arXiv preprint arXiv:1902.04166, 2019.
17. Nguyen T, Le H, Quinn T P, et al. GraphDTA: predicting drug–target binding affinity with graph neural networks[J]. Bioinformatics, 2021, 37(8): 1140-1147.
18. Kipf T N, Welling M. Semi-supervised classification with graph convolutional networks[J]. arXiv preprint arXiv:1609.02907, 2016.
19. Veličković P, Cucurull G, Casanova A, et al. Graph attention networks[J]. arXiv preprint arXiv:1710.10903, 2017.
20. Xu K, Hu W, Leskovec J, et al. How powerful are graph neural networks?[J]. arXiv preprint arXiv:1810.00826, 2018.
21. Yang Z, Zhong W, Zhao L, et al. Mgraphdta: deep multiscale graph neural network for explainable drug–target binding affinity prediction[J]. Chemical science, 2022, 13(3): 816- 833.
22. Karimi M, Wu D, Wang Z, et al. DeepAffinity: interpretable deep learning of compound– protein affinity through unified recurrent and convolutional neural networks[J]. Bioinformatics, 2019, 35(18): 3329-3338.
23. Davis M I, Hunt J P, Herrgard S, et al. Comprehensive analysis of kinase inhibitor selectivity[J]. Nature biotechnology, 2011, 29(11): 1046-1051.10
24. Tang J, Szwajda A, Shakyawar S, et al. Making sense of large-scale kinase inhibitor bioactivity data sets: a comparative and integrative analysis[J]. Journal of Chemical Information and Modeling, 2014, 54(3): 735-743.
25. Guermeur, Yann, et al. "Improved performance in protein secondary structure prediction by inhomogeneous score combination." Bioinformatics (Oxford, England) 15.5 (1999): 413- 421.
26. Combet, Christophe, et al. "NPS@: network protein sequence analysis." Trends in biochemical sciences 25.3 (2000): 147-150.
27. Wang H, Tang J, Ding Y, et al. Exploring associations of non-coding RNAs in human diseases via three-matrix factorization with hypergraph-regular terms on center kernel alignment[J]. Briefings in Bioinformatics, 2021, 22(5): bbaa409.
28. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features[J]. Biopolymers: Original Research on Biomolecules, 1983, 22(12): 2577-2637.
29. Wan L, Zeiler M, Zhang S, et al. Regularization of neural networks using dropcon nect[C]//International conference on machine learning. PMLR, 2013: 1058-1066.
30. Kingma D P, Ba J. Adam: A method for stochastic optimization[J]. arXiv preprint arXiv:1412.6980, 2014.

31. Nair V, Hinton G E. Rectified linear units improve restricted boltzmann machines[C]//Proceedings of the 27th international conference on machine learning (ICML-10). 2010: 807- 814.
32. Zhao Q, Xiao F, Yang M, et al. AttentionDTA: prediction of drug–target binding affinity using attention model[C]//2019 IEEE international conference on bioinformatics and biomedicine (BIBM). IEEE, 2019: 64-69.