Boosting Drug-Target Binding Affinity Predictions with a Novel Three-Branch Convolutional Neural Network Approach

Yaoyao Lu [1] and Hongjie Wu [1]

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

Abstract. The process of discovering new drugs is costly and time-consuming, with safety concerns often arising. Deep learning has become a mainstream approach in computer-aided drug design, with convolutional neural networks (CNN) and graph neural networks (GNN) playing a significant role in drug-target affinity (DTA) prediction. This paper introduces a novel method for predicting DTA using a combination of graph convolutional networks and a three-branch multiscale CNN, leading to significant improvements in prediction accuracy.

Keywords: Drug-Target Binding Affinity Predictions.

1 Introduction

Proteins are involved in all aspects of cellular life activities and play a crucial role in human immunity. The ability to accurately predict drug-target binding affinity is a key focus in the discovery and repositioning of new drugs. Traditional experimental methods have evolved but are limited by being time-consuming and labor-intensive. Computer-aided drug design methods have been developed to save time and labor costs effectively.

Proteins involve all aspects of cellular life activities, and they play a vital role in human immunity [1]. Many diseases are caused by the biochemical dysfunction of protein allogeneic. Specific drugs can alter the way native proteins in the body work, resulting in the desired therapeutic effect [2]. In the discovery and repositioning of new drugs, the ability to accurately predict the drug–target binding affinity becomes the focus of research [3].While experimental methods in wet labs have evolved to screen and characterize chemical molecules, large-scale identification of potential compounds is time-consuming and labor-intensive [4].

In order to save time costs and labor costs, and to make efficient use of resources, many methods of computer-aided drug design have been developed [5]. Virtual screening is one of the main methods. It involves the prediction of potential drugs by many computational models to screen out the drug candidate ligands of interest receptor proteins from large-scale compound ligand libraries. Virtual screening can greatly reduce the number of candidate ligands, significantly reduce the experimental cycle, and thus

accelerate drug discovery [6]. Virtual screening methods can be divided into two categories: receptor-based virtual screening and ligand-based virtual screening methods. Receptor-based virtual screening mainly studies the three-dimensional structure of proteins and seeks interaction with small molecule compounds from the three-dimensional structure [7-9], so it is also called structure-based virtual screening. Common structure-based virtual screening, such as molecular docking [10, 11] and molecular dynamics simulations [12], has been extensively studied.

Although these methods are highly explanatory, their practical application is limited, because they rely heavily on the high-quality threedimensional structure of proteins, and are computationally expensive and inefficient. Ligand-based virtual screening usually starts from the ligand, analyzes the molecular structure and activity information of the known inhibitor, and summarizes the structural characteristics that have an important contribution to the binding ability of the compound by induction. This learned knowledge is then used to screen new ligands to find the compound molecules that meet the requirements [13].

Virtual screening methods are usually based on predictingdrug–target interactions or DTA. The main manifestation is that the input is a vector or graph after the drugs and proteins are encoded, and the output is a classification problem or a regression problem. However, the interactions can be understood as a series of consecutive values used to express the strength of the differentdrug–target interactions. Previously, there were quite a few research ideas that measured drug–target interactions as binary classification tasks [14-18]. In this paper, we focus on DTA prediction. In recent years, deep learning methods have shown excellent performance in many fields [19, 20], and researchers have proposed various data-driven methods based on deep learning [21-24] to study drug targeted binding [25-29].

For example, the deep learning-based DTA prediction model DeepDTA [30], uses a simplified molecular input line entry specification (SMILES) as a drug signature and a protein amino acid as a protein signature. Two features are input into two convolutional neural networks (CNNs) for extraction, and a regression module is then used for prediction through a fully connected layer. GANsDTA [31] is based on a semisupervised generative adversarial network (GAN), which consists of two parts, two GANs for feature extraction and one regression network for prediction. WideDTA [32] takes into account chemical and biological information, using deep learning from four CNNs to predict DTA. DeepAffinity [33] feeds sequences and protein structural properties with drugs into recurrent neural networks (RNNs) for learning.

Deep learning excels in the DTA prediction space [34] and has achieved many achievements. However, in deep learning models [35-37], most experiments express drugs in the form of strings, and the form of one-dimensional sequences is not the natural way molecules are expressed. When we use strings, the structure information of the numerator is lost. The use of graph convolutional networks has also been shown to be more beneficial for computational drug discovery. PADME uses molecular map convolution to predict drug-target interactions, which suggests the potential of GNN in drug development [38]. GraphDTA [39] applied the graph to small molecules to build predictive DTA models for the first time and showed good performance. Although both

CNN-based and graph neural network (GNN)-based approaches have shown good performance at

DTA predictions, there are still some problems that have not been well addressed. First of all, most deep learning methods have only a few CNN layers, and after stacking through convolutional layers, the entire feature information is compressed into a small part, but some local features of the original data are lost. Second, simply using a graph convolutional network (GCN) [9, 39] to graphically express features, does not take into account that the characteristics of each node have different effects on their adjacent and farther nodes, and the closer the node, the greater the impact. To solve the above problems, we propose a method based on the combination of GCN and CNN, putting SMILES into the GCN in the graph, considering the neighboring node weights, and using the attention mechanism based on the GCN. Global and local signatures of proteins are obtained using a three-branched multiscale convolutional neural network (MCNN) [40] at the same time, after which molecular and protein signatures are fused and fed into the prediction module. The prediction module contains three fully connected layers that finally output DTA values.

2 Method

Our approach involves constructing drug molecules into graph representation vectors and learning feature expressions through graph attention networks (GAT) and graph convolutional networks (GCN). A three-branch CNN learns the local and global features of protein sequences, and the two feature representations are merged into a regression module to predict DTA.

In this study, we used one model to deal with drug molecules and another to deal with protein data, for the regression problem of DTA prediction. First, we process the SMILES of drug molecules into graph form with RDKit [41]. The GNN starts with a GAT layer that takes the graph as input and then passes a convolutional feature matrix to the subsequent GCN layers. Each layer is activated by a rectified linear unit (ReLU) function. The final graph representation vector is then computed by concatenating the global max pooling layer and global average pooling layer output by the GCN layer. We represent proteins with amino acid sequences, encode the protein sequences and input them into the embedding layer and then into our CNN. Here, we use a three-branch CNN to extract the local and global signatures of protein amino acid sequences. The three branches use CNNs with different layers and extract different ranges of protein features, which we named the local branch, the middle branch, and the global branch, respectively. After passing through a max pooling layer, the outputs are combined as a protein representation vector. Finally, the molecular representation vector and the protein representation vector are combined and input into the regression module. We use three fully connected layers and set a dropout layer and a ReLU layer after each fully connected layer, and finally output the predicted value of DTA.

2.1 Framework

The framework includes two separate models for drug molecules and protein data. Drug SMILES strings are converted into graph form, and features are extracted using GAT and GCN layers. For proteins, amino acid sequences are encoded and input into a three-branch CNN to extract local and global features. The final molecular and protein vectors are combined and input into the regression module for DTA prediction.

2.2 Model

We evaluate our model on the Davis and KIBA datasets, which are widely used benchmarks for protein and drug binding affinity predictions.

Drugs are represented using SMILES strings converted into graph format RDKit, and proteins are encoded using a 25-tag system based on amino acid properties.

The GAT layer applies a shared linear transformation and calculates attention coefficients for each node. The GCN processes the graph structure, and a multiscale CNN extracts features from protein sequences.

We use the consistency index (CI) and mean squared error (MSE) as metrics to evaluate model performance.

We evaluate our model on two DTA datasets: Davis [42] and KIBA [43]. These two datasets are widely used as benchmark datasets for protein and drug binding affinity predictions. The Davis dataset contains data for selective analysis of kinase protein families and related inhibitors, using dissociation constant (Kd) values [44]. The KIBA dataset combines Kd, the inhibition constant (Ki) [45], or the semi-maximum inhibitory concentration (IC50) [46], using the KIBA value as an affinity. Table 1 summarizes the statistics for both datasets.

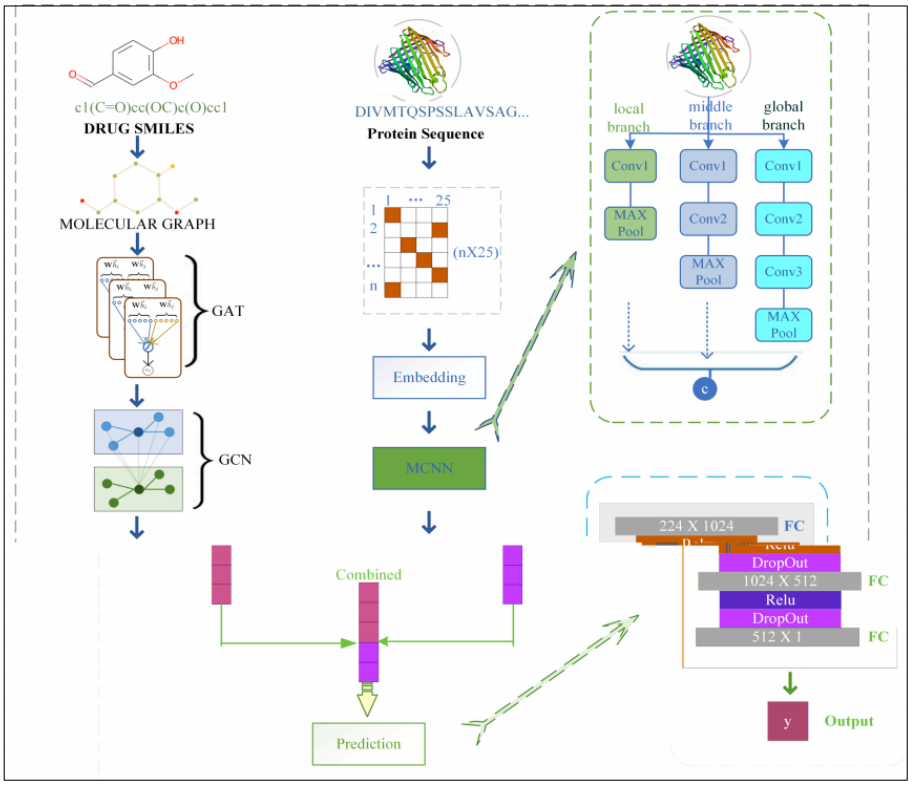
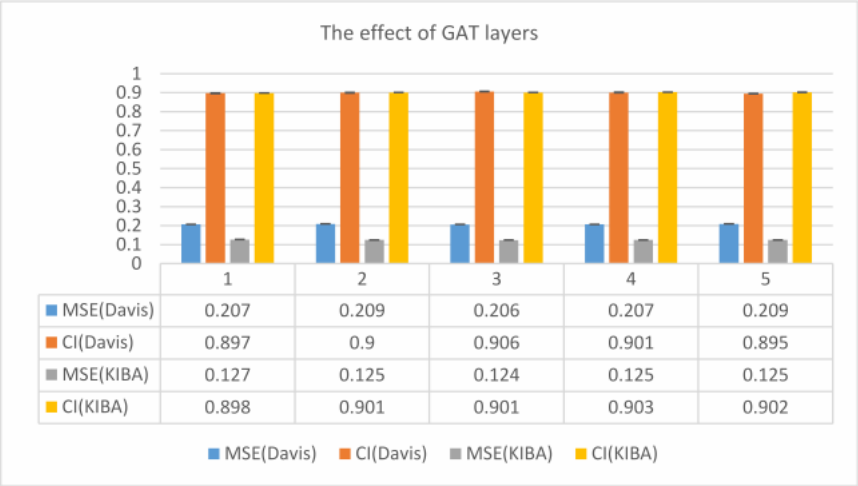


Fig. 1. Architecture.

3 Results

Our model demonstrated a 2.5% improvement in CI and a 21% increase in accuracy as measured by MSE on the Davis dataset compared to DeepDTA. It also outperformed other models including GANsDTA, WideDTA, GraphDTA, and DeepAffinity.



	1	2	3	4	5
MSE(Davis)	0.207	0.209	0.206	0.207	0.209
CI(Davis)	0.897	0.9	0.906	0.901	0.895
MSE(KIBA)	0.127	0.125	0.124	0.125	0.125
CI(KIBA)	0.898	0.901	0.901	0.903	0.902

Fig. 2. Results.

The integration of multiscale CNNs and graph representations for drug molecules and protein sequences, respectively, yielded superior results in DTA prediction. Our model offers a promising approach for accelerating drug discovery.

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] Cao L, Coventry B, Goreshnik I, *et al.* Design of protein-binding proteins from the target structure alone. Nature 2022; 605(7910): 551-60. http://dx.doi.org/10.1038/s41586-022-04654-9 PMID: 35332283
[2] Gonzalez MW, Kann MG. Chapter 4: Protein interactions and disease. PLOS Comput Biol 2012; 8(12): e1002819. http://dx.doi.org/10.1371/journal.pcbi.1002819 PMID: 23300410
[3] Yu JL, Dai QQ, Li GB. Deep learning in target prediction and drug repositioning: Recent advances and challenges. Drug Discov Today 2021; 1359-6446. PMID: 34718208
[4] Deng L, Zeng Y, Liu H, Liu Z, Liu X. DeepMHADTA: Prediction of drug-target binding affinity using multi-head self-attention and convolutional neural network. Curr Issues Mol Biol 2022; 44(5): 2287-99. http://dx.doi.org/10.3390/cimb44050155 PMID: 35678684
[5] Aminpour M, Montemagno C, Tuszynski JA. An overview of molecular modeling for drug discovery with specific illustrative examples of applications. Molecules 2019; 24(9): 1693. http://dx.doi.org/10.3390/molecules24091693 PMID: 31052253
[6] Scior T, Bender A, Tresadern G, *et al.* Recognizing pitfalls in virtual screening: A critical review. J Chem Inf Model 2012; 52(4): 867-81. http://dx.doi.org/10.1021/ci200528d PMID: 22435959
[7] Damale MG, Patil RB, Ansari SA, *et al.* Molecular docking, pharmacophore based virtual screening and molecular dynamics studies towards the identification of potential leads for the management of *H. pylori*. RSC Adv 2019; 9(45): 26176-208. http://dx.doi.org/10.1039/C9RA03281A PMID: 35531003

[8] Loo JSE, Emtage AL, Murali L, Lee SS, Kueh ALW, Alexander SPH. Ligand discrimination during virtual screening of the CB1 cannabinoid receptor crystal structures following cross-docking and
microsecond molecular dynamics simulations. RSC Adv 2019; 9(28): 15949-56. http://dx.doi.org/10.1039/C9RA01095E PMID: 35521393
[9] Jana S, Ganeshpurkar A, Singh SK. Multiple 3D-QSAR modeling, e-pharmacophore, molecular docking, and *in vitro* study to explore novel AChE inhibitors. RSC Adv 2018; 8(69): 39477-95. http://dx.doi.org/10.1039/C8RA08198K PMID: 35558010
[10] Stanzione F, Giangreco I, Cole JC. Use of molecular docking computational tools in drug discovery. Prog Med Chem 2021; 60: 273- 343. http://dx.doi.org/10.1016/bs.pmch.2021.01.004 PMID: 34147204
[11] Rajasekhar S, Karuppasamy R, Chanda K. Exploration of potential inhibitors for tuberculosis *via* structure-based drug design, molecular docking, and molecular dynamics simulation studies. J Comput
Chem 2021; 42(24): 1736-49. http://dx.doi.org/10.1002/jcc.26712 PMID: 34216033
[12] Salo-Ahen OMH, Alanko I, Bhadane R, *et al.* Molecular dynamics simulations in drug discovery and pharmaceutical development. Processes 2020; 9(1): 71. http://dx.doi.org/10.3390/pr9010071
[13] Singh P, Mishra M, Agarwal S, Sau S, Iyer AK, Kashaw SK. Exploring the role of water molecules in the ligand binding domain of PDE4B and PDE4D: Virtual screening based molecular docking of some active scaffolds. Curr Computeraided Drug Des 2019; 15(4): 334-66. http://dx.doi.org/10.2174/1573409914666181105153543 PMID: 30394213
[14] Lim J, Ryu S, Park K, Choe YJ, Ham J, Kim WY. Predicting drug– target interaction using a novel graph neural network with 3D structure-embedded graph representation. J Chem Inf Model 2019; 59(9): 3981-8. http://dx.doi.org/10.1021/acs.jcim.9b00387 PMID: 31443612
[15] Peng J, Wang Y, Guan J, *et al.* An end-to-end heterogeneous graph representation learning-based framework for drug–target interaction prediction. Brief Bioinform 2021; 22(5): bbaa430. http://dx.doi.org/10.1093/bib/bbaa430 PMID: 33517357
[16] Shin B, Park S, Kang K, *et al.* Self-attention based molecule representation for predicting drug-target interaction. arXiv:190806760 2019.
[17] Huang K, Xiao C, Glass LM, Sun J. MolTrans: Molecular interaction transformer for drug–target interaction prediction. Bioinformatics 2021; 37(6): 830-6. http://dx.doi.org/10.1093/bioinformatics/btaa880 PMID: 33070179
[18] Zhao T, Hu Y, Valsdottir LR, Zang T, Peng J. Identifying drug– target interactions based on graph convolutional network and deep neural network. Brief Bioinform 2021; 22(2): 2141-50. http://dx.doi.org/10.1093/bib/bbaa044 PMID: 32367110
[19] Zhang Q, He Y, Wang S, *et al.* Base-resolution prediction of transcription factor binding signals by a deep learning framework. PLoS Comput Biol 2022; 18(3): e1009941. http://dx.doi.org/10.1101/2021.11.01.466840
[20] Shen Z, Zhang Q, Han K, Huang DS. A deep learning model for RNA-protein binding preference prediction based on hierarchical LSTM and attention network. IEEE/ACM Trans Comput Biol Bioinformatics 2022; 19(2): 753-62. PMID: 32750884
[21] Yuan L, Huang DS. A network-guided association mapping approach from DNA methylation to disease. Sci Rep 2019; 9(1): 5601. http://dx.doi.org/10.1038/s41598-019-42010-6 PMID: 30944378
[22] He Y, Shen Z, Zhang Q, Wang S, Huang DS. A survey on deep learning in DNA/RNA motif mining. Brief Bioinform 2021; 22(4): bbaa229. http://dx.doi.org/10.1093/bib/bbaa229 PMID: 33005921
[23] Wang L, You ZH, Huang YA, Huang DS, Chan KCC. An efficient approach based on multi-sources information to predict circRNA – disease associations using deep convolutional neural network. Bioinformatics 2020; 36(13): 4038-46. http://dx.doi.org/10.1093/bioinformatics/btz825 PMID: 31793982

[24] Wang L, You ZH, Huang DS, Zhou F. Combining high speed ELM learning with a deep convolutional neural network feature encoding for predicting protein-RNA interactions. IEEE/ACM Trans Comput Biol Bioinformatics 2020; 17(3): 972-80. http://dx.doi.org/10.1109/TCBB.2018.2874267 PMID: 30296240
[25] Abbasi K, Razzaghi P, Poso A, Ghanbari-Ara S, Masoudi-Nejad A. Deep learning in drug target interaction prediction: current and future perspectives. Curr Med Chem 2021; 28(11): 2100-13. http://dx.doi.org/10.2174/1875533XMTA5qNzU62 PMID: 32895036
[26] Cherkasov A, Muratov EN, Fourches D, *et al.* QSAR modeling: Where have you been? Where are you going to? J Med Chem 2014; 57(12): 4977-5010. http://dx.doi.org/10.1021/jm4004285 PMID: 24351051
[27] Zhang S, Golbraikh A, Tropsha A. Development of quantitative structure-binding affinity relationship models based on novel geometrical chemical descriptors of the protein-ligand interfaces. J Med Chem 2006; 49(9): 2713-24. http://dx.doi.org/10.1021/jm050260x PMID: 16640331
[28] Politi R, Rusyn I, Tropsha A. Prediction of binding affinity and efficacy of thyroid hormone receptor ligands using QSAR and structure-based modeling methods. Toxicol Appl Pharmacol 2014; 280(1): 177-89. http://dx.doi.org/10.1016/j.taap.2014.07.009 PMID: 25058446
[29] Wang S, Jiang M, Zhang S, *et al.* Mcn-cpi: Multiscale convolutional network for compound–protein interaction prediction. Biomolecules 2021; 11(8): 1119. http://dx.doi.org/10.3390/biom11081119 PMID: 34439785
[30] Öztürk H, Özgür A, Ozkirimli E. DeepDTA: Deep drug–target binding affinity prediction. Bioinformatics 2018; 34(17): i821-9. http://dx.doi.org/10.1093/bioinformatics/bty593 PMID: 30423097
[31] Zhao L, Wang J, Pang L, Liu Y, Zhang J. GANsDTA: Predicting drug-target binding affinity using GANs. Front Genet 2020; 10: 1243. http://dx.doi.org/10.3389/fgene.2019.01243 PMID: 31993067
[32] Öztürk H, Ozkirimli E, Özgür A. WideDTA: prediction of drugtarget binding affinity. arXiv:190204166 2019.
[33] 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; 35(18): 3329-38. http://dx.doi.org/10.1093/bioinformatics/btz111 PMID: 30768156
[34] Mayr A, Klambauer G, Unterthiner T, *et al.* Large-scale comparison of machine learning methods for drug target prediction on ChEMBL. Chem Sci 2018; 9(24): 5441-51. http://dx.doi.org/10.1039/C8SC00148K PMID: 30155234
[35] Yi HC, You ZH, Huang DS, Li X, Jiang TH, Li LP. A deep learning framework for robust and accurate prediction of ncRNAprotein interactions using evolutionary information. Mol Ther Nucleic Acids 2018; 11: 337-44. http://dx.doi.org/10.1016/j.omtn.2018.03.001 PMID: 29858068
[36] Chuai G, Ma H, Yan J, *et al.* DeepCRISPR: optimized CRISPR guide RNA design by deep learning. Genome Biol 2018; 19(1): 80. http://dx.doi.org/10.1186/s13059-018-1459-4 PMID: 29945655
[37] Shen Z, Zhang YH, Han K, *et al.* miRNA-disease association prediction with collaborative matrix factorization. Biomolecular Networks for Complex Diseases 2017; 2017
[38] Feng Q, Dueva E, Cherkasov A, *et al.* Padme: A deep learningbased framework for drug-target interaction prediction. arXiv:180709741 2018.
[39] Nguyen T, Le H, Quinn TP, Nguyen T, Le TD, Venkatesh S. GraphDTA: Predicting drug–target binding affinity with graph neural networks. Bioinformatics 2021; 37(8): 1140-7. http://dx.doi.org/10.1093/bioinformatics/btaa921 PMID: 33119053

[40] Yang Z, Zhong W, Zhao L, Yu-Chian CC. MGraphDTA: deep multiscale graph neual network for explainable drug–target binding affinity prediction. Chem Sci 2022; 13(3): 816-33. http://dx.doi.org/10.1039/D1SC05180F PMID: 35173947
[41] Bento AP, Hersey A, Félix E, *et al.* An open source chemical structure curation pipeline using RDKit. J Cheminform 2020; 12(1): 51. http://dx.doi.org/10.1186/s13321-020-00456-1 PMID: 33431044
[42] Davis MI, Hunt JP, Herrgard S, *et al.* Comprehensive analysis of kinase inhibitor selectivity. Nat Biotechnol 2011; 29(11): 1046-51. http://dx.doi.org/10.1038/nbt.1990 PMID: 22037378