Prediction And Analysis of Drug-Target Interaction Via Comparative Learning on Heterogeneous Networks

Jiarui Mou *

Yingcai Honors College, University of Electronic Science and Technology of China, Chengdu, China

* Corresponding Author Email: 2024150901011@std.uestc.edu.cn

Abstract. In this paper, the rapid development of drug target interaction in the field of Computer Science in recent years is analyzed in depth, and Heterogeneous Network-based Contrastive Learning for Drug–Target Interaction (HNCL-DTI) is used as the anchor point for developmental exploration, and its data set limitations and model limitations are expanded, in order to integrate and sort out the technical application and development level in this field, and strive to find better integrated applications that improve and integrate scattered technologies in this field. Drug target interaction prediction is an important research direction of bioinformatics. Traditional methods often ignore the edge characteristics when dealing with heterogeneous biomedical networks, resulting in insufficient prediction accuracy. Some feasible paths to improve the recent methods are discussed. With its limitations and corresponding supplements, it is possible to explore better technical models in the field of drug target interaction, which is expected to help better drug target models in the future.

Keywords: Prediction of drug target interaction; Heterogeneous graph neural network; Heterogeneous biomedical networks; Figure attention; Comparative learning.

1. Introduction

Drug–Target Interaction (DTI) field has been inseparable from computer science disciplines such as in-depth learning, and thanks to the development of computer field and the gradual deepening of interdisciplinary, the application of cross fusion in drug target interaction prediction has become an important direction in a sense. The mainstream methods adopted in this direction in recent years can be roughly divided into three branches. First, from the perspective of data, DTI has introduced heterogeneous graph attention from the single network of Drug–Target Interaction Network (DTINet) to the multimodal multi view of Knowledge Graph-enhanced Drug–Target Interaction prediction (KG-DTI) and Unified interacton representation via co-embedding on heterogeneous networks (Unicorn) in Heterogeneous Graph Attention Network (HGAT) [1-4]. Second, from the network structure, it can be noted that in this field, from the shallow Graph Neural Network (GNN) of HGAT to the three-layer GNN of Subgraph Graph Contrastive Learning for Drug–Target Interaction (SGCL-DTI), then the hierarchical Transformer of Hierarchical Multi-level Self-Attention for Drug–Target Interaction (Hi-MulSA-DTI) and Global Convolutional Former (GCFormer) [5]. Third, from the use of learning paradigm, from the random walk of DTINet, no learning, to the self-monitoring model of self-monitoring/comparative learning Contrastive Predictive Coding (CPC) and Simple Contrastive Learning of visual Representations (SimCLR), to SGCL-DTI and Heterogeneous Network-based Contrastive Learning for Drug–Target Interaction (HNCL-DTI), the application of comparative learning has been strengthened, and DrugLAMP has made a contribution to the loss of integration and comparison within Transformer [6]. Although the field of drug target interaction prediction has made important development in recent years, the overall structure is still loose, and the cohesion and logical level of each part can be further sorted out and optimized. The integration and sorting can be conducive to further research and model improvement by subsequent researchers.

Drug target interaction prediction is the core issue in modern drug discovery and development. This paper believes that DTINet has opened the door to DTI prediction, and HGAT is its important

supplement and expansion. Heterogeneous graphs are introduced into DTI prediction for the first time. Among them, DTI prediction using heterogeneous neural network (HGNN) is considered to be a very promising method, in which the attention-based model usually shows excellent performance [2]. However, existing methods usually ignore the edge characteristics when dealing with heterogeneous biomedical networks, which may lead to the inaccuracy of prediction results [6]. HNCL-DTI (a comparative learning method based on heterogeneous networks to predict drug target interactions), as a recent innovation and repeatedly verified method, first introduced heterogeneous comparative learning into the field of drug target prediction, and introduced the edge attention mechanism to cooperate with the node attention. This method technically integrates three core components: heterogeneous graph neural network (HGAT), node/edge dual attention and comparative learning. It can be said that it is a master in the field of DTI prediction based on machine learning. At the same time, the new method DTI-RME has also attracted wide attention, but it has not been widely verified by many parties. For the sake of the scientificity and preciseness of the content, this method is not mentioned in this paper.

In the second chapter, this paper will introduce the mainstream research methods and the characteristics of mainstream data sets in turn. In Chapter 3, the paper will discuss the performance of the mainstream method HNCL-DTI on its corresponding dataset HBN-A/HBN-B. Secondly, in Chapter 4, the limitation analysis of mainstream data sets including BindingDB, DrugBank, KEGG-DTI and HBN-A/HBN-B will also be clearly summarized and explained, including common industry problems such as cold start and rare data volume. In Chapter 5, some limitations of today's model will be pointed out. Finally, the feasibility analysis of the scheme will be strengthened by transplanting the recent achievements in the general field.

2. Research Methods and Data

2.1. Core Technology Architecture

The core technology architecture of DTINet can be summarized as four modules: multi network integration, restart of random walk feature diffusion, low rank matrix approximate reduction and supervised classification. The design goal is to compress more than 10 heterogeneous biological networks into one-dimensional feature vectors of drug target pairs, and then use a simple classifier to complete DTI prediction and drug relocation, so as to realize DTI prediction.

Later, HGAT integrated three modules, namely, heterogeneous graph convolution, double-layer attention and task output layer, and designed heterogeneous graphs with different node types, edge relationships and semantic paths. The focus is on how to distinguish node types and automatically assign weights to neighbors and meta paths.

Then, the core technology architecture of SGCL-DTI, which deepens GNN to three layers, can be understood as five steps: first coding heterogeneous graphs, then building drug protein pair (DPP) nodes, then applying topology semantic dual view, then conducting supervised collaborative comparative learning, and finally end-to-end two classification. Its innovation is to change the form of comparative learning from self-supervised learning to supervised learning, which is no longer randomly augmented. Instead, it uses the known DTI tags to pull the same DPP nodes closer and push the different ones away on the two same DPP graphs of topology and semantics, so as to maximize the use of limited positive samples.

As a master of the application of machine learning and comparative learning in DTI field, the core of HNCL-DTI method is the organic combination of three key components. One is heterogeneous graph neural network, which is a kind of graph neural network architecture that uses type differentiation transformation and attention aggregation to learn unified node representation on a graph containing multiple types of nodes and edges. The second is the dual attention mechanism, which includes node attention (finding the most important one in a group of elements) and edge attention (identifying the key connection from the connection relationship). The third is comparative learning. By constructing positive sample pairs (similar samples) and negative sample

pairs (dissimilar samples), the training model is also similar to SGCL-DTI to narrow the positive samples and push away the negative samples in the feature space, so as to learn the general feature representation of data.

2.2. Dataset Description

BindingDB this dataset contains about 2.3M DTI records measured experimentally, and its data contains affinity values. The affinity continuous value can be used to train the regression model. The DrugBank dataset contains about 20 K drugs and about 5 K proteins, which are friendly to cold start and have complete drug properties (SMILES, targets, pharmacological categories). The KEGG-DTI data set correlates the three tables of pathway drug target, which is more suitable for heterogeneous mapping, and has its own pathway label, which means that it can be used for pathway enrichment interpretation. HNCL-DTI research uses two standard datasets, of which HBN-A is the benchmark dataset (11655 data points) introduced in DTINet, which is widely used in DTI prediction tasks involving heterogeneous biological data. It is a classic dataset and can be regarded as the threshold of new models in this field [3]. Another data set HBN-B (15322 data points) is the extended version of HBN-A, which is generated by HGAN. The data is larger and the test is more challenging, which can be regarded as the final standard for judging the feasibility of the model. In addition, the HNCL-DTI data set HBN-A/HBN-B integrates six public databases including DrugBank, KEGG, UniProt, GO, SIDER/CTD (common side effects database) and DisGeNet/OMIM (disease gene database), including six types of relationships (D-T, D-D, D-I, D-S, T-T, T-I), which are authoritative and reliable.

2.3. Evaluation Index

This field includes six evaluation indicators: AUC, AUPR, Precision, Recall, F1 and MCC. Where Precision is the number of positive samples in which the prediction is positive. Recall is how many real positive samples have been successfully found. F1 is the result of harmonizing and averaging Precision and Recall. MCC takes TP, TN, FP and FN into account and gives a comprehensive correlation coefficient from -1 to+1. TP is the model saying yes and true, TN is the model saying no and true no, FP is the model saying yes but false, FN is the model saying no but false. These four numbers are MCC from negative one to positive one. The closer it is to+1, the more accurate it is. If the result is+1, it means that the prediction is completely consistent with the truth. Where AUC is the probability that the model will rank the positive sample ahead of the negative sample when one positive sample and one negative sample are randomly selected. At the same time, AUPR focuses on the area of the accuracy check full curve of positive samples, which is more stable for unbalanced data.

3. Experimental Results and Performance Evaluation of HNCL-DTI

3.1. Benchmark Comparison

HNCL-DTI was compared with nine baseline methods: DTINet [1], NeoDTI, MultiDTI, IMCHGAN, SGCL-DTI (a supervised drug protein joint characterization learning) [5], HGAN, RGCN, HGT, Simple-HGN.

3.2. Performance

Among them, HNCL-DTI ranked first in all six indicators on the HBN-A dataset, which was significantly improved compared with the previous best baseline. The model successfully predicted six real DTIs that had been verified in the literature but were unknown at the time of training. On the HBN-B data set, the overall performance was the best, with four first and two second. In the prediction, the predicted targets of DB00228 and DB00408 were 100% consistent with the latest research. In terms of specific data, HNCL-DTI has pushed the AUC to 0.956 (HBN-A) and 0.929

(HBN-B), and the AUPR reached 0.961/0.927, and maintained a leading position under the 10 folds cross validation+cold start split, 2-3 percentage points higher than the best model in 2023. Although there is a gap from the new method DTI-RME, it is still the real mainstream and most commonly used method, with the most repetition and more reliability.

Here, this paper believes that this method has basically reached the highest level at the time of publication, with high prediction accuracy, which was the mainstream and commonly used method in this field at that time.

4. Analysis of Data Set Limitations

4.1. Main Limitations

BindingDB, DrugBank, KEGG-DTI and HBN-A/HBN-B have their own limitations. There may be multiple records of the same drug protein pair in the BindingDB data set, which need to be de duplicated and averaged to reduce the error. Due to the long history of the DrugBank data set, some drugs have been delisted, so the data set has great era limitations. KEGG-DTI dataset with a small amount of data may be more suitable for small sample or migration scenarios, but not for large sample scenarios.

The data set limitations faced by HBN-A/HBN-B are mainly reflected in the data size constraints, data quality problems and uneven annotation.

The data volume of HBN-A and HBN-B is relatively small, which may reflect the inherent difficulties of data collection in this field. According to the author's understanding, the construction of DTI data sets in this industry requires a lot of professional knowledge and experimental verification, resulting in the scarcity of high-quality annotation data. Therefore, DrugLAMP proposes a shallow message passing framework mode with semantic structure dual channels, which can integrate the multimodal features of drugs or proteins with very little data or parameters, and can still achieve high accuracy and migrate to predict the drug target interaction even in the extreme scenario of zero samples [7]. Unfortunately, HNCL-DTI does not use or compare DrugLAMP as the basis. Therefore, it can be considered that the HNCL-DTI method is likely to be affected by the data scale, and this factor needs to be taken into account in its ability to process information and data. Because HNCL-DTI method relies on known DTI annotation, it is conservative and has the risk of incomplete annotation or error. The existing DTI prediction methods are still troubled by the high sparsity of DTI data sets and the cold start problem [3], and the known drug target interactions account for a small proportion relative to all possible combinations. Finally, in the process of multi-source database integration, there are annotation noise problems such as homonymy, synonymy and so on.

The above four data sets show that the DTI prediction field generally has the problems of unfriendly cold start, small amount of data and no support of big data. This problem may be one of the factors that lead to the accuracy of drug target interaction still need to be improved. If the data set can solve this industry pain points in the future, it will promote the development of this field.

4.2. HNCL-DTI's Solutions to These Problems

In view of the above problems, this paper found that HNCL-DTI adopted the following solutions, and sorted out the solutions to some industry problems in this paper as Table 1:

Table 1. Three Scheme comparing

Question type	Solution	Effect
Small data scale	Integrate six public databases to build heterogeneous graphs, expand the number of nodes to 11000-15000, and expand the number of edges to>100000	Effective expansion
Annotation sparsity and cold start	Pseudo tags are constructed by multi hop path similarity+network topology consistency. The positive sample coverage is about 5%, and the positive negative ratio is reduced from 1:100 to 1:10	Significant improvement
Uneven marking quality	PubChem CID and UniProt AC are used as the primary key to remove duplicate and secondary correction of conflict labels	Quality improvement
Heterogeneous features are difficult to align	Uniformly map to the same node attribute table, align first and then enter the graph	Avoid subsequent uncertainty

5. Analysis on The Limitations of Domain Model Centered on HNCL-DTI Model

5.1. Technical Limitations

First, the innovative nature of HNCL-DTI is questionable. The dual attention mechanism may only be a simple addition of the two methods calculated separately, not a real innovation. Previous HAN has proposed heterogeneous graph attention network. The main contribution of HNCL-DTI is actually to apply it to DTI prediction, which is suspected of simple transplantation [2].

Second, the comparative learning method may be too basic. Although the HNCL-DTI method uses the method of difficult negative sampling, it is not based on the difficult negative sampling of biological similarity. There may be a risk of error caused by ignoring the biological similarity of drug target pairs, and the application of InfoNCE loss function does not give a specific biological customized design. In addition, CPC in 2018 gave the general public InfoNCE of comparative learning. In 2020, simclr made the formula engineering in the field of vision to the extreme, and did not make corresponding adjustments to the biological design, so it remains doubtful [8, 9].

Third, the HNCL-DTI fusion strategy is simple, but the simple fusion strategy $H = \frac{H^{\text{node}} + H^{\text{edge}}}{2}$ adopted may cause serious information loss. In contrast, HMSA-DTI has proposed a hierarchical multimodal self-attention mechanism, which can capture the interaction within and between modes at the same time, which is significantly better than simple feature stitching or averaging, and HMSA-DTI uses more advanced cross attention and hierarchical fusion. In contrast, the method of HNCL-DTI is outdated at least at the technical level, so it is more appropriate to adopt the hierarchical multimodal self-attention mechanism of HMSA-DTI in this regard [10].

Fourth, the model has an implicit assumption vulnerability. HNCL-DTI directly calculates the cosine similarity between node features and edge features and sends them to InfoNCE. It is assumed that they are aligned in the same metric space, but there is no proof in the text, which may lack certain theoretical support.

5.2. Main Sources of Performance Improvement

Despite the above limitations, the performance improvement of HNCL-DTI mainly comes from the explicit modeling of edge semantics, cross view consistency constraints and domain verification

closed loop. The explicit modeling of edge semantics takes the relational semantics of six types of edges as a learnable vector for the first time in the Han framework, which significantly enhances the expression ability of complex pharmacological relationships. Cross view consistency constraint uses InfoNCE loss function to compare the representations of node view and edge view, and align different modal information at the unsupervised level [8, 9]. Finally, the end-to-end experimental verification is completed on the HBN-A/b constructed by the public biological database. The case proves that the improvement of DTI prediction by the above two points is brought about by the model structure, rather than the data scale or parameter adjustment bonus.

6. Feasibility Analysis of Improvement Scheme

Through the analysis and comparison of the technologies proposed in the top-level conference of machine learning (NeurIPS, ICML, ICLR) from 2024 to 2025, it is found that the following improvement schemes may be able to migrate to the drug target field. The first is difficult negative sampling. The difficult negative sampling strategy is trained by specially selecting difficult to distinguish negative samples, which should be better than random selection in theory [11]. The latest research shows that the method of creating positive and negative sample sequences through the hybrid token generator and extracting valuable information using the comparative learning framework can significantly improve the quality of node representation learning, providing a new technical path for the difficult negative sampling in DTI prediction [11]. The second is gradient balanced contrastive learning, which mainly solves the gradient conflict between contrastive learning and the main task. The unicorn framework proposed by Feng et al. effectively integrates multi view molecular representation learning through a unified comparative learning method, providing a new technical path for solving the gradient conflict problem in DTI prediction [4]. This method has achieved the most advanced performance in the molecular representation learning task, and is expected to lay an expanded foundation for improving the generalization ability and prediction accuracy of DTI prediction model to some extent. Finally, hierarchical attention, compared with the traditional single-mode method, recent multimodal methods such as Hierarchical Multimodal Self-Attention-based graph neural network for Drug–Target Interaction prediction significantly improve the prediction performance by integrating multiple molecular representations [10].

The advantages of these general technologies lie in their high degree of automation and advanced technology. In theory, they should be able to be directly applied and transplanted into the HNCL-DTI framework, and it is expected to improve the model to a certain extent.

7. Conclusion

This paper briefly describes the influential methods of drug target mutual prediction based on machine learning, and finds that the mainstream method HNCL-DTI performs well on its corresponding data set HBN-A/HBN-B, which is the most advanced and comprehensive method in its publication period. The mainstream datasets include BindingDB, DrugBank, KEGG-DTI and HBN-A/HBN-B. There are many problems in DTI prediction field, such as unfriendly cold start, small amount of data and no support of big data. And in the analysis of the limitations of the domain model centered on HNCL-DTI model, the performance of the method is still affected by the questionable nature of innovation, the possibility that the comparative learning method may be too basic, the simple fusion strategy, and the existence of implicit hypothesis loopholes. Finally, the feasibility analysis of the scheme is strengthened by the transplantation of recent achievements in the general field.

DTINet introduced a basic framework for mutual prediction of drug targets, while HGAT introduced heterogeneous graphs for the first time, and HNCL-DTI is of great significance in technology integration and application innovation. By integrating heterogeneous graph neural

network, dual attention mechanism and comparative learning, this method has achieved excellent performance on the data set HBN-A/HBN-B.

However, by comparing with the research results of peers in the same period or even in the early stage, this method still has limitations in data set size, model architecture innovation and technology implementation. The sparsity and size limitation of data sets reflect the general challenges in the interdisciplinary field of medicine and computer, and the composition of the model may be more reflected in the patchwork and integrated application of existing technologies, reflecting the lag of the migration and application of computer science methods in the field of DTI.

Throughout the DTI prediction field, it is not difficult to find an interesting technology timeline. DTINet has opened up the situation of heterogeneous network applications and become the originator of other methods. In the field of drug target interaction prediction, it has gradually changed from random walk to the path of self-monitoring and comparative learning such as HNCL-DTI, as well as the layered Transformer of GCFormer DTI on the network structure, and the data perspective from a single network tends to be multimodal and multi-view. On the whole, the introduction of specific methods and technologies has been added, and the interdisciplinary is becoming more and more profound.

Through the comparative analysis with the latest technologies in the general field, it is found that the advanced technologies such as difficult negative sampling, adaptive comparative learning and hierarchical attention have high transplantation feasibility, which is expected to further improve the prediction performance of HNCL-DTI. In general, this paper summarizes the main research progress and achievements in the field of drug target interaction prediction in recent years, and also provides a technical path with certain value. At the same time, it also provides possible improvement directions and some theoretically feasible paths for subsequent research in this field.

References

[1] Luo Y, Zhao X, Zhou J. et al. A network integration approach for drug-target interaction prediction and computational drug repositioning from heterogeneous information. Nat Commun 8, 573, 2017. https://doi.org/10.1038/s41467-017-00680-8

[2] Wang X, Ji H, Shi C, Wang B, Ye Y, Cui P, Yu PS. Heterogeneous graph attention network. In Proceedings of the 2019 World Wide Web Conference, 2022–2032, 2019.

[3] Gao M, Zhang D, Chen Y, Li Z, Li X. A unified drug–target interaction prediction framework based on knowledge graph and recommendation system. Brief Bioinform 24(1), bbac578, 2023.

[4] Li Z, Wang Y, Liu H, Zhang X. UniCorn: A unified contrastive learning approach for multi-view molecular representation learning. Adv Neural Inform Process Syst 35, 1–14, 2022.

[5] Li Y, Qiao G, Gao X, Wang G. Supervised graph co-contrastive learning for drug–target interaction prediction. Bioinformatics 38(10), 2847–2854, 2022. https://doi.org/10.1093/bioinformatics/btac164

[6] Hu J, Bewong M, Kvashie S, Zhang W, Nofong VM, Wu G, Feng Z. A heterogeneous network-based contrastive learning approach for predicting drug-target interaction. arXiv, 2024.

[7] Chen Q, Zhang X, Li Y, Wang X, Zhu H, Li J. Accurate and transferable drug–target interaction prediction with DrugLAMP. Bioinformatics 40(12), btae693, 2024. https://doi.org/10.1093/bioinformatics/btae693

[8] Van den Oord A, Li Y, Vinyals O. Representation learning with contrastive predictive coding. Adv Neural Inform Process Syst 32, 6926–6936, 2019.

[9] Chen T, Kornblith S, Norouzi M, Hinton G. A simple framework for contrastive learning of visual representations. In Proceedings of the 37th International Conference on Machine Learning, 1597–1607, 2020.

[10] Bian J, Lu H, Dong G, Wang G. Hierarchical multimodal self-attention-based graph neural network for DTI prediction. Brief Bioinform 25(4), bbae293, 2024.

[11] Chen H, Li X, Zhang Y, Wang J. Leveraging contrastive learning for enhanced node representations in tokenized graph transformers. In Proceedings of the 12th International Conference on Learning Representations, 2024.