Sparse Imbalanced Drug-Target Interaction Prediction via Heterogeneous Data Augmentation and Node Similarity

Runze Wang[1], Zehua Zhang[1(✉)], Yueqin Zhang[1], Zhongyuan Jiang[2], Shilin Sun[1], and Chenwei Zhang[3]

[1] Taiyuan University of Technology, Taiyuan 030024, China
wangrunze0317@link.tyut.edu.cn, {zhangzehua,zhangyueqin}@tyut.edu.cn
[2] Xidian University, Xian 710068, China
zyjiang@xidian.edu.cn
[3] Amazon, Seattle, WA 98109, USA
cwzhang@amazon.com

Abstract. Drug-Target Interaction (DTI) prediction usually devotes to accurately identify the potential binding targets on proteins so as to guide the drug development. However, the sparse imbalance of known drug-target pairs remains a challenge for high-quality representation learning of drugs and targets, interfering with accurate prediction. The labeled drug-target pairs are far less than the missed since the obtained DTIs are recorded with pathogenic proteins and sophisticated bio-experiments. Therefore, we propose a deep learning paradigm via **H**eterogeneous graph data **A**ugmentation and node **S**imilarity (**HAS**) to solve the sparse imbalanced problem on drug-target interaction prediction. Heterogeneous graph data augmentation is devised to generate multi-view augmented graphs through a heterogeneous neighbors sampling strategy. Then the consistency across different graph structures is captured using graph contrastive optimization. Node similarity is calculated on the heterogeneous entity association matrices, aiming to integrate similarity information and heterogeneous attribute gain for drug-target interaction prediction. Extensive experiments show that HAS offers superior performance in sparse imbalanced scenarios compared state-of-the-art methods. Ablation studies prove the effectiveness of heterogeneous graph data augmentation and node similarity.

Keywords: Sparse imbalanced DTI prediction · Heterogeneous graph data augmentation · Graph contrastive optimization · Node similarity

1 Introduction

Drug-Target Interaction prediction plays an essential role in the drug discovery process [1]. And it often leads to the next stages of pharmacological in vitro experiments [2]. The growing clinical demands pose the challenges to drug

J. Gama et al. (Eds.): PAKDD 2022, LNAI 13280, pp. 548–561, 2022.
https://doi.org/10.1007/978-3-031-05933-9_43

screening based on traditional experiments. The emergence of machine learning has brought a new boom in computer-aided drug design which reduces the time-consuming and expensive bio-experiments [3]. Some computational approaches for DTI prediction were proposed in supervised learning view, such as applying deep learning techniques to extract chemical features from known structure data [4–6] or analyze the potential correlation among labeled drug-target pairs [7,8]. Several studies attempted to perform the semi-supervised tasks with known and unknown drug-target pairs, including modeling the tripartite relations of drug-target-disease [9], constructing the heterogeneous information networks and leveraging the diverse biological entity properties to alleviate the negative impact of missed DTI labels [10,11]. Although the efforts have been made on respective tasks, the supervised learning methods rely on the both chemical structure data and labels, and the semi-supervised learning methods are based on the hypothesis of balanced positive and negative samples (i.e., the known drug-target pairs are treated as positive samples, while the unknown interacting pairs are regarded as negative samples), neglecting the realistic issue that positive samples are far less than negative samples. Drug discovery usually builds on the pathogenic proteins [12]. Pharmaceutical researchers screen the candidate drugs that change the proteins bioactivity. Only if the changes of proteins bioactivity meet the clinical and research needs, can DTIs be recorded. Large number of DTIs are missed and the obtainable DTI labels are limited by the amount of discovered pathogenic proteins. Furthermore, the real-world drug-target interactions far exceed the recorded, causing the observed DTIs are extra sparse compared with the whole drug-target pairs space. The sparse imbalanced interacting drug-target pairs are insufficient to learn high-quality feature representations for drugs and targets which leads to inaccurate prediction.

To sum up, we propose a deep learning paradigm by integrating **H**eterogeneous Data **A**ugmentation and Node **S**imilarity for sparse imbalanced DTI prediction, named as **HAS**. Especially, we present a heterogeneous graph data augmentation module to generate multi-view augmented graphs on constructed heterogeneous graph involving the node types of drug and target. Differentiate from the recent studies of contrastive learning on homogeneous graph [13,14], a heterogeneous contrastive learning strategy is designed to capture the agreement between different graph structures. Based on the general assumption that drugs with higher similarity are more likely to have common linked target [3], node similarity is calculated on the heterogeneous entity associated matrices. So far, the intrinsic topological structure information and node similarity information from different attribute spaces are acquired to supplement the sparse supervised signal. The main works are summarized as follows:

- We formalize the sparse imbalance problem on drug-target interaction prediction, and present a novel deep learning method via heterogeneous graph data augmentation and node similarity to solve.
- Heterogeneous graph data augmentation is designed to capture intrinsic and universal structure patterns between multi-view augmented graphs. The

similarity information and heterogeneous attribute information are incorporated to strengthen the features of drugs and targets.
– Empirical studies on the real-world datasets demonstrate that HAS has significant improvement in sparse imbalanced DTIs scenario compared with the state-of-the-art methods.

2 Related Work

DTI prediction has attracted much attention in recent years. Numerous studies are dedicated to reducing the search space of drug candidates and facilitating drug discovery process [15]. And the related methods can be mainly divided into three aspects: bio-feature extraction, pairwise similarity discovery and bioinformatics network mining.

Bio-feature extraction takes chemical structure data as the input of deep learning framework to extract the main features of drugs and targets respectively, and finally fusing the features of both to predict DTI. For example, the works DeepDTA [4] designs deep learning models to predict the binding affinity (one type of DTI) using sequential data of drugs and proteins. DeepConv-DTI [5] ensembles local residue patterns of proteins. Graph neural network (GNN) is reported as a powerful tool in graph embedding tasks [16], a computational approaches named GraphDTA [6] is proposed to capture molecular topological features of drugs with GNN to improve the prediction performance. Such methods inevitably rely on known drug-target pairs and structure data.

Pairwise similarity discovery mainly measures the similarity between multiple drug-target pairs, which is used as the interaction information. MATT_DTI [17] introduce multi-head attention mechanism to obtain the similarity information of different drug-target pairs. Chen *et al.* [8] present to utilize a transformer-based decoder that extract interaction features substructure pairs of drugs and proteins. These methods have great effort by incorporating the similarity information into interaction prediction, but the complex network relations are unconsidered, e.g. drug-drug.

Bioinformatics network mining aims at using graph representation learning methods to predict drug-target interactions on the heterogeneous network. The work NeoDTI [10] constructs heterogeneous network with drug, target, disease, etc. and predict drug-target interactions in graph reconstruction way. MultiDTI [11] maps all the heterogeneous biological entities to common feature space, so the space distances between nodes are regarded as prediction scores of DTI. EEG_DTI [18] applies graph neural network to learn embedding vectors of drugs and targets for DTI prediction. However, these studies make the number of positive and negative samples approximate the balanced, ignoring the realistic problem that the known drug-target interactions are sparse in the whole drug-target pairs space. HAS focuses on the sparse imbalanced DTI prediction that belongs to an urgent real-world issue. Heterogeneous graph data augmentation and node similarity are proposed from topology-level and node-level to alleviate the negative impact brought by sparse known drug-target pairs.

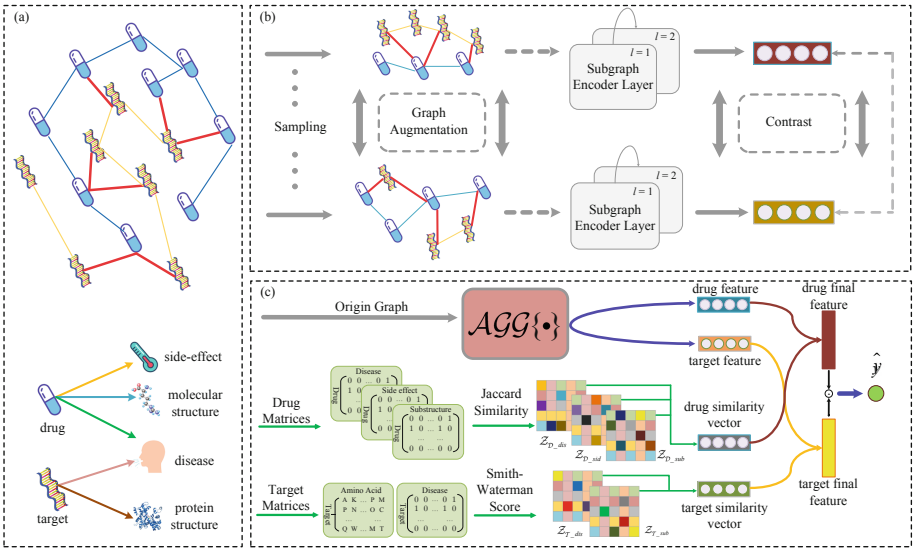


Fig. 1. Illustration of the proposed HAS. (a) The upper half is heterogeneous information network including drugs and targets. The bolded red edges represent the drug-target interactions, the blue and yellow edges show drug-drug interactions and target-target interactions, respectively. The lower half depicts heterogeneous relations about drug and target. (b) Heterogeneous graph data augmentation module first generates multi-view augmented graphs through heterogeneous neighbors sampling, then encodes subgraph structure and learns the nodes features, finally maximizes the agreement of same node from different views via contrastive learning. (c) Node similarity information is calculated on the heterogeneous associated matrices. The learned features of drugs and targets on the origin graph are learned using the aggregate function. Next two types of features are fused as final feature representation to predict the DTI probability. (color figure online)

3 Sparse Imbalanced DTI Prediction

The final goal is to predict interactions between drugs and targets, so heterogeneous information network with only drug and target nodes is defined as $\mathcal{HG} = \{\mathcal{D},\mathcal{T},\mathcal{E},\mathcal{R}\}$, where $\mathcal{D}$ and $\mathcal{T}$ denote sets of drugs and targets. $\mathcal{E}$ and $\mathcal{R}$ are sets of edges and edge types, which are associated with relational matrices, drug-drug matrix $\mathcal{M}_{\mathcal{D}\_\mathcal{D}}$, drug-target matrix $\mathcal{M}_{\mathcal{D}\_\mathcal{T}}$, target-target matrix $\mathcal{M}_{\mathcal{T}\_\mathcal{T}}$. For matrix element $m_{(i,j)} \in \{0,1\}$, if $m_{(i,j)} = 1$, existing $e_{(i,j)} \in \mathcal{E}$.

Given the heterogeneous graph $\mathcal{HG}$, the known edges between drugs and targets are far less than the unknown drug-target interaction edges since only DTIs meeting the clinical needs will be recorded. The final DTI prediction can be cast as an edge classification task via learning the prediction function $\mathcal{F}\{\cdot\}$ under the sparse imbalanced condition, which is formulated as follows:

$$\widehat{y} = \mathcal{F}\{(d_i, t_j), \mathcal{HG} | d_i \in \mathcal{D}, t_j \in \mathcal{T}, \mathcal{E}^{+}_{\mathcal{D}\_\mathcal{T}} << \mathcal{E}^{-}_{\mathcal{D}\_\mathcal{T}}\} \tag{1}$$

where $\widehat{y}$ represents the predicted interaction probability between drug d_i and t_j, $\mathcal{E}^{+}_{\mathcal{D}\_\mathcal{T}}$ denotes set of known DTIs and $\mathcal{E}^{-}_{\mathcal{D}\_\mathcal{T}}$ is set of unknown DTIs. The symbol '$\ll$' indicates the $|\mathcal{E}^{+}_{\mathcal{D}\_\mathcal{T}}|$ is much less than $|\mathcal{E}^{-}_{\mathcal{D}\_\mathcal{T}}|$.

4 Heterogeneous Graph Data Augmentation and Node Similarity

The framework of HAS is shown in Fig. 1. Heterogeneous graph data augmentation module adopts the graph contrastive learning to capture the intrinsic graph structure pattern from different augmented views. Node similarity module is devised to incorporate similarity information between nodes and heterogeneous attribute information for DTI prediction.

4.1 Heterogeneous Graph Data Augmentation

Mining the inherent pattern of heterogeneous graph suitably is beneficial for its representation learning.

Multi-view Graph Augmentation. Different from the recent works of graph contrastive learning that build generators on homogeneous graph, HAS focuses on generating augmented graphs on the heterogeneous graph including various node types. Besides, the imbalanced distribution of multi-typed edges causes the number of neighboring nodes varies from each node. Overall, the multi-view generator is designed through a heterogeneous neighbors sampling strategy, the sampled drugs and targets are derived by random walk with restart. This way of augmented graph generation can avoid the imbalanced problem that edges with heterogeneous types and establishing message propagation with high-order nodes as far as possible. The implementation process takes drug nodes as example:

1. Taking current drug node d_0 as starting point of random walk with restart, the iterative walk is performed to its neighboring node which is either drug or target, and the next step could be itself with probability π. The walk will stop until set Γ_{d_0} about d_0 successfully collects nodes with fixed number.
2. According to the node set Γ_{d_0}, walking path and their related edges on the original graph, a random heterogeneous subgraph $\mathcal{G}_{d_0}$ is generated. $\mathcal{G}_{d_0}$ is regarded as an augmented version in a view with the core node d_0. Repeat the above process twice to obtain two augmented graphs $\mathcal{G}^{(1)}_{d_0}$, $\mathcal{G}^{(2)}_{d_0}$.

Similarly, if the target node t_0 is used as the 'hub' node, the generated augmented graphs are denotes as $\mathcal{G}^{(1)}_{t_0}$, $\mathcal{G}^{(2)}_{t_0}$.

Heterogeneous Subgraph Encoding. The researches of heterogeneous graph learning [19,20] analyze the inherent heterogeneity that the features of different types of nodes may fall in different feature space. In this paper, we consider

that drugs and targets are heterogeneous on the sampled subgraphs, a heterogeneous graph neural network is adopted to aggregate the neighboring attribute with different types. Since different types of nodes contribute differently to its embedding, and so do the different nodes with the same type, we employ attention mechanism in GNN layers to weight the aggregated neighbors messages for each node. First the embeddings of nodes are initialized. Then for each drug node d_i, the attention coefficients are calculated with its neighboring drug nodes:

$$\alpha_{(d_i,d_j)}^{(l)} = \frac{\exp\{\text{LeakyReLU}(\mathbf{a}_{\mathcal{D}}^{(l)^T}[\mathbf{h}_i^{(l)} \oplus \mathbf{h}_j^{(l)}])\}}{\sum_{k\in\mathcal{N}_{\mathcal{D}}(i)} \exp\{\text{LeakyReLU}(\mathbf{a}_{\mathcal{D}}^{(l)^T}[\mathbf{h}_i^{(l)} \oplus \mathbf{h}_k^{(l)}])\}} \tag{2}$$

where $\alpha_{(d_i,d_j)}^{(l)}$ is the attention coefficient between drug d_i and it neighboring drug d_j, l denotes the current layer of heterogeneous graph neural network, LeakyReLU$(\cdot)$ is the nonlinear activation function, $\mathbf{h}_i^{(l)}$ and $\mathbf{h}_j^{(l)}$ represent the hidden feature vectors of d_i and d_j at $l-$th layer. $\mathbf{a}_{\mathcal{D}}^{T}$ is transposed attention vector between drug nodes, $\oplus$ defines the concatenation of two vectors and $\mathcal{N}_{\mathcal{D}}(i)$ defines the set for d_i with drug type neighbors. If the neighbors are target nodes, the heterogeneous attention scores are computed as follows:

$$\beta_{(d_i,t_j)}^{(l)} = \frac{\exp\{\text{LeakyReLU}(c_{\mathcal{D}}^{(l)^T}[\mathbf{h}_i^{(l)} \oplus \mathcal{W}_{\mathcal{D}}^{(l)}\mathbf{p}_j^{(l)}])\}}{\sum_{k\in\mathcal{N}_{\mathcal{T}}(i)} \exp\{\text{LeakyReLU}(c_{\mathcal{D}}^{(l)^T}[\mathbf{h}_i^{(l)} \oplus \mathcal{W}_{\mathcal{D}}^{(l)}\mathbf{p}_k^{(l)}])\}} \tag{3}$$

where $\beta_{(d_i,t_j)}^{(l)}$ defines the computed heterogeneous attention score between drug d_i and t_j at $l-$th layer. $c_{\mathcal{D}}^{T}$ and $\mathcal{N}_{\mathcal{T}}(i)$ denote the transposed attention vector and the set for d_i with target type neighbors, respectively. $\mathbf{p}_j^{(l)}$ and $\mathcal{W}_{\mathcal{D}}^{(l)}$ are the learned hidden feature of target and the feature mapping matrix from target space to drug space. Finally, the feature aggregation in a grouping way is performed to update the 'hub' drug node feature according calculated homogeneous and heterogeneous attention coefficients:

$$\mathbf{h}_i^{(l+1)} = \text{ReLU}((\sum\nolimits_{d_j\in\mathcal{N}_{\mathcal{D}}(i)} \alpha_{(d_i,d_j)}^{(l)}\mathbf{h}_j^{(l)} + \sum\nolimits_{t_j\in\mathcal{N}_{\mathcal{T}}(i)} \beta_{(d_i,t_j)}^{(l)}\mathcal{W}_{\mathcal{D}}^{(l)}\mathbf{p}_j^{(l)})\mathcal{W}^{(l)} + \mathbf{b}^{(l)}) \tag{4}$$

where ReLU$(\cdot)$ is a nonlinear activation function, $\mathcal{W}^{(l)}$ and $\mathbf{b}^{(l)}$ define the learnable feature transformation matrix and bias vector. After L-layer graph neural network, the drug feature representation $\mathbf{h}_i^{(1)}$of d_i is obtained in subgraph $\mathcal{G}_{d_0}^{(1)}$, as well as $\mathbf{h}_i^{(2)}$ in $\mathcal{G}_{d_0}^{(2)}$. Analogously, the learned representation of any target node t_i is calculated in two views of data augmentation as $\mathbf{p}_i^{(1)}$, $\mathbf{p}_i^{(2)}$.

Graph Contrastive Optimization. By this, the nodes features of drugs and targets are learned containing the multi-view subgraph structure information. The recent studies use contrastive learning [21] to optimize the self-supervised learning task that maximize the agreement between positive samples. In order to

discover the universal graph topological feature between two augmented graphs, we devise the optimizer in a graph contrastive learning manner. The features under different views of the same node are defined as the positive pairs and the features under different views of different nodes are defined as the negative pairs. For example, $(\mathbf{h}_i^{(1)}, \mathbf{h}_i^{(2)})$ of drug d_i is regarded as positive pair, $(\mathbf{h}_i^{(1)}, \mathbf{h}_j^{(2)})$ of drug d_i and d_j is regarded as negative pair. Then, the contrastive loss $\mathcal{L}_{\mathcal{D}}$ related to drug is calculated as follows:

$$\mathcal{L}_{\mathcal{D}} = \sum\nolimits_{d_i \in \mathcal{D}} -\log \frac{\exp(\text{sim}(\mathbf{h}_i^{(1)}, \mathbf{h}_i^{(2)})/\tau)}{\sum_{d_j \in \mathcal{D}, i \neq j} \exp(\text{sim}(\mathbf{h}_i^{(1)}, \mathbf{h}_j^{(2)})/\tau)} \tag{5}$$

where $\text{sim}(\cdot)$ is the cosine similarity function and τ defines the temperature parameter. Similarly, the contrastive loss about target can be obtained as $\mathcal{L}_{\mathcal{T}}$.

4.2 Node Similarity

Based on the general assumption that drugs with high similarity may share common interactions with the same target, we incorporate drug-drug, target-target similarity information to enrich the feature of drugs and targets. The direct associated biological entities can be viewed as heterogeneous attributes, so we calculate the node similarity on the associated matrices. The chemical structures of drugs are comprised of SMILES strings, a cheminformatics tool named RDKit is used to convert SMILES strings to morgan fingerprints that are expressed as binary vectors. Each entry demonstrates the presence or absence of certain chemical substructure. Then the substructure feature matrix $\mathcal{M}_{\mathcal{D}\_sub}$ of all drugs can be acquired. Given the biological association matrices $\mathcal{M}_{\mathcal{D}\_sid}$ for drug-side effect, $\mathcal{M}_{\mathcal{D}\_dis}$ for drug-disease, $\mathcal{M}_{\mathcal{T}\_dis}$ for target-disease, and drug substructure matrix $\mathcal{M}_{\mathcal{D}\_sub}$, the principal components analysis algorithm is employed to tackle the negligible vector sparsity and high-dimensional issues. Next the similarities of drug-drug or target-target are calculated by the Jaccard similarity measure. After that, similarity matrices from different heterogeneous attribute spaces can be obtained: $\mathcal{Z}_{\mathcal{D}\_sid}$ in side effect space, $\mathcal{Z}_{\mathcal{D}\_dis}$ in disease space, $\mathcal{Z}_{\mathcal{D}\_sub}$ in substructure space and a target similarity matrix $\mathcal{Z}_{\mathcal{T}\_dis}$ in disease space. The protein structure consists of amino acids sequence. Considering the co-occurrence of local functional fragments in different protein, we choose the Smith-Waterman score measure as the similarity calculation means between proteins. The protein substructure similarity matrix is denoted as $\mathcal{Z}_{\mathcal{T}\_sub}$. Finally, the respective similarity matrices are fused:

$$\mathcal{Z}_{sim}^{\mathcal{D}} = \mathcal{Z}_{\mathcal{D}\_sid} \oplus \mathcal{Z}_{\mathcal{D}\_dis} \oplus \mathcal{Z}_{\mathcal{D}\_sub}, \quad \mathcal{Z}_{sim}^{\mathcal{T}} = \mathcal{Z}_{\mathcal{T}\_dis} \oplus \mathcal{Z}_{\mathcal{T}\_sub} \tag{6}$$

$\mathcal{Z}_{sim}^{\mathcal{D}}$ and $\mathcal{Z}_{sim}^{\mathcal{T}}$ are the fused similarity matrices of drug and target and the row vector contains the similarity and heterogeneous attribute information.

4.3 DTI Prediction Task and Optimization

Here we aim to perform the DTI prediction on the original graph $\mathcal{HG}$. Duo to the existing heterogeneity on $\mathcal{HG}$, a node feature aggregation function with attention

Algorithm 1. Sparse Imbalanced DTI prediction based on HAS.

Input: Graph $\mathcal{HG}$= $\{\mathcal{D}, \mathcal{T}, \mathcal{E}, \mathcal{R}\}$, Matrices $\mathcal{M}_{\mathcal{D}\_sid}$,$\mathcal{M}_{\mathcal{D}\_dis}$,$\mathcal{M}_{\mathcal{D}\_sub}$,$\mathcal{M}_{\mathcal{T}\_dis}$,$\mathcal{M}_{\mathcal{T}\_sub}$
Output: Predicted drug-target interaction probability $\widehat{y}$

1: Generate multi-view augmented graphs $\mathcal{G}_{d_i}^{(1)}$, $\mathcal{G}_{d_i}^{(2)}$, $\mathcal{G}_{t_j}^{(1)}$, $\mathcal{G}_{t_j}^{(2)}$
2: Heterogeneous attention subgraph encoding using Equation (2)(3)(4)
3: Maximize the agreement of positive pairs from different views using Equation (5)
4: Get node similarity information and heterogeneous attribute information via similarity computing on the matrices $\mathcal{M}_{\mathcal{D}\_sid}$, $\mathcal{M}_{\mathcal{D}\_dis}$, $\mathcal{M}_{\mathcal{D}\_sub}$, $\mathcal{M}_{\mathcal{T}\_dis}$, $\mathcal{M}_{\mathcal{T}\_sub}$
5: Apply weighted aggregation function (7) and feature fusing function (8) to learn node embeddings on $\mathcal{HG}$
6: Predict the interaction probability using Equation (9)

weights is applied to feature learning on augmented graphs similarly:

$$\mathbf{h}_i^{\mathcal{HG}} = \mathcal{AGG}_{j\in\mathcal{N}_{\mathcal{D}}(i),k\in\mathcal{N}_{\mathcal{T}}(j)}^{\mathcal{HG}}\{\mathbf{h}_j^{\mathcal{HG}}, \mathbf{p}_k^{\mathcal{HG}}, \alpha_{(d_i,d_j)}^{\mathcal{HG}}, \beta_{(d_i,t_k)}^{\mathcal{HG}}\} \tag{7}$$

where $\mathcal{AGG}\{\cdot\}$ denotes the weighted node aggregation function, $\alpha_{(d_i,d_j)}^{\mathcal{HG}}$ is the attention score between drug d_i and drug d_j on the original graph, $\beta_{(d_i,t_k)}^{\mathcal{HG}}$ is the attention score between drug d_i and target t_k. Analogously, the target nodes features on the original graph can be acquired using function $\mathcal{AGG}\{\cdot\}$. For the purpose of taking full advantage of known drug-target interaction information and similarity information, we utilize a multi-layer fusion function to fuse the learned nodes features on $\mathcal{HG}$ and the computed similarity features:

$$\mathbf{h}_i^{final} = \mathcal{FC}_{\Theta}(\mathbf{h}_i^{\mathcal{HG}} \oplus \mathbf{z}_i^{\mathcal{D}}), \quad \mathbf{p}_j^{final} = \mathcal{FC}_{\Theta}(\mathbf{p}_j^{\mathcal{HG}} \oplus \mathbf{z}_j^{\mathcal{T}}) \tag{8}$$

where $\mathbf{h}_i^{final}$ and $\mathbf{p}_j^{final}$ denote the final features of drug d_i and target t_j. $\mathcal{FC}_{\Theta}(\cdot)$ is the multi-layer fusion function and Θ is set of trainable parameters. $\mathbf{z}_i^{\mathcal{D}}$ and $\mathbf{z}_j^{\mathcal{T}}$ represent the similarity vectors of drug d_i and target t_j. The final layer predicts the probability via calculating the inner product of vectors:

$$\widehat{y}_{(d_i,t_j)} = \text{Sigmoid}(\mathbf{h}_i^{final} \odot \mathbf{p}_j^{final}) \tag{9}$$

where $\widehat{y}_{(d_i,t_j)}$ denotes the predicted probability between drug d_i and target t_j, $\odot$ and Sigmoid$(\cdot)$ represent the dot product measure and sigmoid nonlinear function. As the final DTI prediction task is treated as edge classification, we adopt the cross-entropy loss to fit prediction score and the label value:

$$\mathcal{L}_{pre} = -\sum\nolimits_{(d_i,t_j)\in\mathcal{E}_{\mathcal{D}\_\mathcal{T}}^{+}} \log(\widehat{y}_{(d_i,t_j)}) - \sum\nolimits_{(d_i,t_k)\in\mathcal{E}_{\mathcal{D}\_\mathcal{T}}^{-}} \log(1-\widehat{y}_{(d_i,t_k)}) \tag{10}$$

To complete the whole optimization task that the DTI prediction under sparse imbalance condition, we combine the loss of both data augmentation and DTI prediction together, which is defined as follows:

$$\mathcal{L} = \mathcal{L}_{pre} + \xi_1(\mathcal{L}_{\mathcal{D}} + \mathcal{L}_{\mathcal{T}}) + \xi_2||\Theta||_2^2 \tag{11}$$

$||\Theta||_2^2$ is the L_2-norm term that prevents training overfitting. ξ_1 and ξ_2 are hyperparameters that control the loss of data augmentation and the L_2-norm term. Algorithm 1 shows the DTI prediction procedure of our proposed framework.

5 Experiments

To evaluate the effectiveness of the proposed method and discuss the reasons, we conduct extensive experiments with different sparsity settings.

5.1 Datasets and Experiment Setup

Experiments are conducted on the constructed drug-target network and associated matrices following Luo *et al.* [22], where drug-drug interactions and drug-target interactions are extracted from DrugBank (Version 3), protein-protein interactions are extracted from HPRD database Release 9. Others are that associated disease data from Comparative Toxicogenomics Database, related side effect data from SIDER database Version 2. The SMILES strings for drugs and amino acid sequences for proteins are obtained following Zhou [13]. The details of heterogeneous entities are summarized in Table 1.

Table 1. The statistics of datasets

Entity type	Numbers	Relation type	Numbers	Sparse ratio
Drug	708	Drug-target	1923	0.00179
Target	1512	Drug-drug	10036	
Disease	5603	Drug-disease	199214	
Side effect	4192	Drug-side effect	80164	
		Target-target	7363	
		Target-disease	1596745	

The compared baselines cover the recent state-of-art methods and traditional deep learning-based models, which all perform drug-target interaction prediction on heterogeneous biological networks. DTINet (2017) combines the unsupervised feature learning from heterogeneous biological network and matrix completion for DTI prediction. NeoDTI (2019) tends to train the model by reconstructing the edge on the heterogeneous graph. MultiDTI (2021) maps the biological entities into vector space aiming to minimize the distance between the entities features. In addition, we consider Graph Attention Network (GAT) and Deep Neural Network (DNN) algorithms as contrast group.

Experiments were conducted on Inspur heterogeneous cluster GPU:12 *32 G Tesla V100 s, memory 640 G DDR2. We deploy the HAS framework with Pytorch and DGL. About the training process of model, we use Adam optimizer with the learning rate of 0.005. The dimension of the initialized features is set as 128, the restart probability π and the temperature parameter τ are set as 0.8 and 0.07, respectively. The final task is denoted as edge classification, we evaluate the DTI prediction performance using Area Under the Receiver Operating Characteristic Curve (AUROC) and Area Under the Precision-Recall Curve (AUPRC).

5.2 Results Discussion

Comparison with Baselines in Different Sparse Ratios. We examine the DTI prediction performance of HAS under sparse imbalanced condition. To further explore its robustness, the positive-negative ratio is adjusted to simulate different sparse DTIs scenarios. The 10-fold cross-validation is implemented on all positive samples and randomly negative samples that are selected according to sparse ratio. We split 90% positive and negative samples in each fold dataset for training, 10% for test purposes.

In addition to simulate the realistic issue that the known DTIs are far less than unknown DTIs, we also consider the experimental setup of baselines and design the experiment with balanced positive-negative samples. Table 2 shows the result comparison with baselines. 1:10 is that the negative samples are 10 times to positive samples, 1:all represents all negative samples are used. Particularly, we have the following observations:

Table 2. Performance comparison with baselines in different sparsity setting

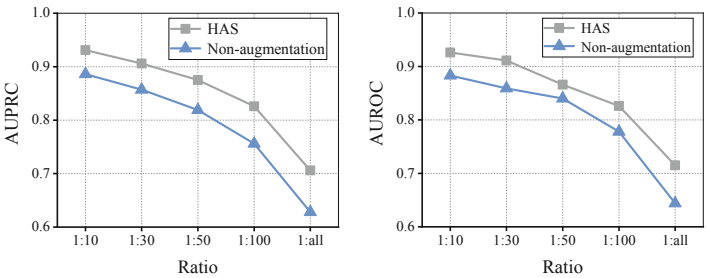
Method	AUPRC						AUROC					
	1:1	1:10	1:30	1:50	1:100	1:all	1:1	1:10	1:30	1:50	1:100	1:all
DNN	0.765	0.691	0.645	0.582	0.441	0.326	0.776	0.755	0.712	0.646	0.597	0.535
GAT	0.873	0.800	0.724	0.612	0.533	0.405	0.825	0.801	0.761	0.723	0.662	0.496
DTINet	0.932	0.865	0.816	0.757	0.671	0.507	0.914	0.873	0.845	0.789	0.692	0.522
NeoDTI	NA	0.874	0.835	0.784	0.726	0.602	NA	**0.943**	0.890	0.839	0.790	0.662
MultiDTI	**0.947**	0.921	0.878	0.837	0.782	0.656	**0.961**	0.891	0.866	0.818	0.730	0.633
HAS	0.938	**0.931**	**0.906**	**0.865**	**0.817**	**0.706**	0.945	0.926	**0.911**	**0.874**	**0.832**	**0.715**
Improv.	NA	1.10%	3.19%	3.35%	4.48%	7.62%	NA	NA	2.36%	4.17%	5.32%	8.01%

(1) The sparse imbalanced interactions between drugs and targets limit efficient prediction performance. We can see that all the models perform well on the balanced DTI prediction. However, with the negative sample increases, the results show a significant decreasing trend. When the negative pairs are sampled to 100 times, model performance drops more than 15%. Until all negative pairs are joined, the metrics drop dramatically again by nearly 15% compared with 1:100 sparse scenario. It confirms the aforementioned statement that a large number of missed drug-target pairs have negative impact on learning high-quality features representation for drugs and targets. Because the rare drug-target interactions cause the weak supervised signals on heterogeneous graph, message propagation between nodes is less to represent graph structure.

(2) HAS expresses the superior improvement. We find that the improvements of HAS mainly come from the sparse imbalanced DTIs scenarios. For example, a 3.35% gain (AUPRC) and 4.17% gain (AUROC) over MultiDTI when the negative pairs are sampled up to 50 times. Furthermore, HAS significantly outperforms alternative approaches by 7.62% (AUPRC) and 8.01% (AUROC).

We conclude that HAS is less affected by negative effect of sparse imbalanced drug-target interactions than the compared baselines. It may be that HAS could capture the intrinsic and universal graph structure feature from topology level as well as similarity information between nodes from node level. All above are used to enhance feature learning when a large amounts of DTIs are missed. MultiDTI achieve the best on the balanced DTI prediction as it adopts a oversampling strategy that oversamples the positive samples by 10 times and under-samples the negative samples. NeoDTI tends to perform DTI prediction under sparse condition, we use 'NA' to label it in Table 2.

Benefits of Heterogeneous Graph Data Augmentation. Heterogeneous graph data augmentation module is proposed to learn the intrinsic graph patterns. We examine the effectiveness of the module in sparse imbalanced DTIs scenarios. The results are shown in Fig. 2(a) and Fig. 2(b), where 'Non-augmentation' means without using heterogeneous graph data augmentation.

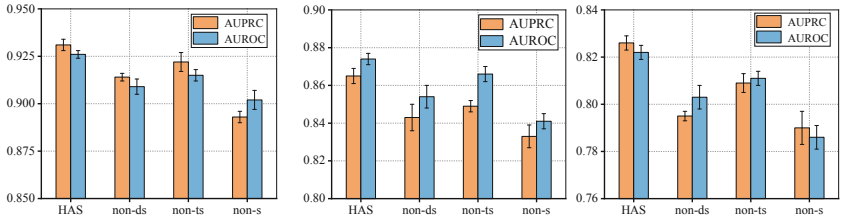


(a) Comparison in AUPRC. (b) Comparison in AUROC.

Fig. 2. Effectiveness of heterogeneous graph data augmentation.

As expected, the prediction performance of the model without data augmentation drops significantly compared to the overall HAS. Specifically, a clear trend is emerging that the sparser the DTIs data is, the better the data augmentation performs. We observe quite significant drop in AUPRC and AUROC if all the negative drug-target pairs are used. It illustrates that the heterogeneous graph data augmentation contributes more to performance improvements under sparse imbalanced DTIs condition. The augmented graphs encompass the multi-view graph structure information and the contrastive learning optimizes the association between augmented graphs to capture the universal graph structure, which can be used to supplement the missed interactions information.

Benefits of Node Similarity. To test the effectiveness of node similarity information with sparse known drug-target pairs, we first simulate three different sparse imbalanced scenarios as shown in Fig. 3(a) (1:10), Fig. 3(b) (1:50) and Fig. 3(c) (1:100). The experiments between node similarity (**HAS**) and non-similarity (**non-s**) are conducted. Besides, only using drug similarity (**non-ds**) and only using target similarity (**non-ts**) are set so as to explore the importance of drug similarity information and target similarity information for DTI prediction. The results indicate average 4% drop (AUPRC) and 3% drop (AUROC) without using node similarity information. This verifies that the joined similarity information provide positive impact with sparse known drug-target pairs. And diverse attribute information from heterogeneous entities can better characterize the latent properties of drugs and targets. In addition, we can find that both drug nodes similarity and target nodes similarity can make a contribution, which can be used to enrich the learned features.



(a) 1:10 (b) 1:50 (c) 1:100

Fig. 3. Effectiveness of node similarity.

Performance Comparison with Different Layers. Considering that GNN with various layers have differences in node feature learning, we perform multi-combination of heterogeneous layers in order to seek the most beneficial setting for the DTI prediction. The experimental results can be seen in Table 3. The performance of HAS achieve the best if setting 2 layers for augmented graphs and origin graph. The setting of 2 layers on the augmented graphs outperforms the setting of 3 layers. It may be smaller size of node data on subgraphs, the aggregation of 2-hop neighboring nodes covers enough drugs and targets. The stacking of aggregated layers will no longer perform significantly better.

Table 3. HAS performance with different layers.

Augmented graph	Origin graph	AUPRC	AUROC
Layer = 2	Layer = 1	0.880	0.898
	Layer = 2	0.931	0.926
	Layer = 3	0.922	0.919
Layer = 3	Layer = 1	0.911	0.923
	Layer = 2	0.920	0.923
	Layer = 3	0.908	0.917

6 Conclusion

In this work, we formulate the sparse imbalance problem on drug-target interaction prediction and analyze the reason. Especially, we propose a deep learning framework HAS to solve it via heterogeneous graph data augmentation and node similarity. Heterogeneous graph data augmentation pursues to capture the intrinsic graph structure pattern from different augmented versions. Node similarity information is incorporated for DTI prediction. Experimental results show that HAS outperforms the baselines in various sparse imbalanced DTIs scenarios. Ablation studies verify the effectiveness of proposed heterogeneous graph data augmentation and node similarity to alleviate the sparse imbalance issue.

The complicate bio-experiments in drug discovery cause that the real labeled drug data is less accessible. The future work will explore and construct the other heterogeneous biological networks to strengthen generalization of model. And the impact of augmented graphs scale will be further investigated.

Acknowledgements. This work was supported by the National Natural Science Foundation of China (61503273, 61702356), Industry-University Cooperation Education Program of the Ministry of Education, and Shanxi Scholarship Council of China.

References

1. Sun, M., Zhao, S., Gilvary, C.: Graph convolutional networks for computational drug development and discovery. Briefings Bioinform. **21**(3), 919–935 (2020)
2. Vamathevan, J., Clark, D., Czodrowski, P.: Applications of machine learning in drug discovery and development. Nat. Rev. Drug Discov. **18**(6), 463–477 (2019)
3. Bagherian, M., Sabeti, E., Wang, K.: Machine learning approaches and databases for prediction of drug-target interaction: a survey paper. Briefings Bioinf. **22**(1), 247–269 (2021)
4. Hakime, Ö.: DeepDTA: deep drug-target binding affinity prediction. Bioinformatics **34**(17), 821–829 (2018)
5. Lee, I., Keum, J., Nam, H.: DeepConv-DTI: prediction of drug-target interactions via deep learning with convolution on protein sequences. PLoS Comput. Biol. **15**(6), e1007129 (2019)

6. Nguyen, T., Le, H., Quinn, T.P.: GraphDTA: predicting drug-target binding affinity with graph neural networks. Bioinformatics **37**(8), 1140–1147 (2021)
7. Huang, K., Xiao, C., Glass, L.M.: MolTrans: molecular interaction transformer for drug-target interaction prediction. Bioinformatics **37**(6), 830–836 (2021)
8. Chen, L., Tan, X., Wang, D.: TransformerCPI: improving compound-protein interaction prediction by sequence-based deep learning with self-attention mechanism and label reversal experiments. Bioinformatics **36**(16), 4406–4414 (2020)
9. Chen, H., Li, J.: Modeling relational drug-target-disease interactions via tensor factorization with multiple web sources. In: WWW (2019)
10. Wan, F., Hong, L., Xiao, A.: NeoDTI: neural integration of neighbor information from a heterogeneous network for discovering new drug-target interactions. Bioinformatics **35**(1), 104–111 (2019)
11. Zhou, D., Xu, Z., Li, W.T.: MultiDTI: drug-target interaction prediction based on multi-modal representation learning to bridge the gap between new chemical entities and known heterogeneous network. Bioinformatics **37**(23), 4485–4492 (2021)
12. Xia, X.: Bioinformatics and drug discovery. Curr. Top. Med. Chem. **17**(15), 1709–1726 (2017)
13. Qiu, J., Chen, Q., Dong, Y.: Gcc: graph contrastive coding for graph neural network pre-training. In: KDD, pp. 1150–1160 (2020)
14. You, Y., Chen, T., Sui, Y.: Graph contrastive learning with augmentations. In: NeurIPS, pp. 5812–5823 (2020)
15. Jung, L.S., Cho, Y-R.: Survey of network-based approaches of drug-target interaction prediction. In: BIBM, pp. 1793–1796 (2020)
16. Wu, Z., Pan, S., Chen, F.: A comprehensive survey on graph neural networks. IEEE Trans. Neural Netw. Learn. Syst. **32**(1), 4–24 (2020)
17. Zeng, Y., Chen, X., Luo, Y.: Deep drug-target binding affinity prediction with multiple attention blocks. Briefings Bioinform. **22**(5), bbab117 (2021)
18. Peng, J., Wang, Y., Guan, J.: An end-to-end heterogeneous graph representation learning-based framework for drug-target interaction prediction. Briefings Bioinform. **22**(5), bbaa430 (2021)
19. Zhang, C., Song, D., Huang, C.: Heterogeneous graph neural network. In: KDD, pp. 793–803 (2019)
20. Wang, X., Ji, H., Shi, C.: Heterogeneous graph attention network. In: WWW, pp. 2022–2032 (2019)
21. Wu, J., Wang, X., Feng, F.: Self-supervised graph learning for recommendation. In: SIGIR, pp. 726–735 (2021)
22. Luo, Y., Zhao, X., Zhou, J.: A network integration approach for drug-target interaction prediction and computational drug repositioning from heterogeneous information. Nat. Commun. **8**(1), 1–13 (2017)