Supervised Graph Co-contrastive Learning for Drug-Target Interaction Prediction

Yang Li¹, Guanyu Qiao¹, Xin Gao² and Guohua Wang¹ *

¹ College of Information and Computer Engineering, Northeast Forestry University, Harbin, 150004, China.
² King Abdullah University of Science and Technology, Mathematical and Computer Sciences and Engineering, Thuwal, Kingdom of Saudi Arabia

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Identification of Drug-Target Interactions (DTIs) is an essential step in drug discovery and repositioning. DTI prediction based on biological experiments is time-consuming and expensive. In recent years, graph learning based methods have aroused widespread interest and shown certain advantages on this task, where the DTI prediction is often modeled as a binary classification problem of the nodes composed of drug and protein pairs. Nevertheless, in many real applications, labeled data are very limited and expensive to obtain. With only a few thousand labeled data, models could hardly recognize comprehensive patterns of DPP node representations, and are unable to capture enough commonsense knowledge, which is required in DTI prediction. Supervised contrastive learning gives an aligned representation of DPP node representations with the same class label. In embedding space, DPP node representations with the same label are pulled together, and those with different labels are pushed apart.

Results: We propose an end-to-end supervised graph co-contrastive learning model for DTI prediction directly from heterogeneous networks. By contrasting the topology structures and semantic features of the drug-protein-pair network, as well as the new selection strategy of positive and negative samples, SGCL-DTI generates a contrastive loss to guide the model optimization in a supervised manner. Comprehensive experiments on three public datasets demonstrate that our model outperforms the SOTA methods significantly on the task of DTI prediction, especially in the case of cold start. Furthermore, SGCL-DTI provides a new research perspective of contrastive learning for DTI prediction.

Availability: The research shows that this method has certain applicability in the discovery of drugs, the identification of drug-target pairs and so on.

Contact: yli@nefu.edu.cn

Supplementary information: Supplementary data are available at Bioinformatics online.

1 Introduction

Drug targets are proteins that can be targeted by drugs and produce effects in cells. The identification of interactions between drugs and protein targets is not only an essential step in drug discovery (Feng et al., 2017), but also provides guidance towards drug repositioning, multi drug pharmacology, drug resistance (Xue et al., 2018) and side effect prediction (Mongia and Majumdar, 2020). However, it is time consuming and expensive to determine drug-target interactions (DTIs) by biological experiments.

In order to speed up drug discovery, researchers have been exploring computational methods to identify DTIs. Existing methods of drug-target interaction prediction can be classified into three categories: ligand-based (Keiser et al., 2007), docking-based (Shaikh et al., 2016), and machine learning-based approaches (Wen et al., 2017). In addition to the widely used traditional algorithms, a remarkable trend is to approach DTIs from a network perspective. Graph representation learning models such as graph convolutional network (GCN) (Kipf and Welling, 2016) and graph attention network (GAT) (Velikovi et al., 2017) have been used to learn from various types of homogeneous or heterogeneous network information for DTI prediction (Sun et al., 2020). Most existing methods are mainly divided into two independent steps: first to extract the representation vectors of drugs and proteins, and second to apply a deep neural network to predict the final label based on the representation. To the best of our knowledge, there are only few end-to-end models to directly predict the interactions between drugs and proteins from many heterogeneous networks.

In addition to the early computational DTI prediction methods, such as ligand-based (González-Díaz et al., 2011) and docking-based methods (Cheng et al., 2012; Meng et al., 2017), machine learning methods have attracted great attention, as they enable large scale testing of candidates in a relatively short time (Peng et al., 2020; Gao et al., 2018). A key idea behind these methods is the hypothesis of "guilt-by-association", which means that similar drugs may have similar targets and vice versa (Luo et al., 2017). Based on this assumption, machine learning models such as random forest
(RF), decision tree (DT) and support vector machine (SVM) etc., are widely used as classifiers to predict whether a DTI is present or not.

On the other hand, some advanced machine learning methods have been proposed to integrate external information for DTI prediction. Yamanishi et al. (2010) developed a bipartite graph model where the chemical and genomic spaces as well as the drug-protein interaction network are integrated into a pharmacological space. Xia et al. (2010) proposed a semi-supervised learning method, Laplacian regularized least square (LapRLS), to utilize both the small amount of available labeled data and the abundant unlabeled data together in order to give the maximum generalization ability from the chemical and genomic spaces. At the same time, pharmacological or phenotypic information, such as side-effects (Mizutani et al., 2012), transcriptional response data (Iorio et al., 2010), drug–disease associations (Wang et al., 2014), public gene expression data (Sirota et al., 2011) and functional data have been incorporated in DTI to provide diverse information and a multi-view perspective for predicting novel DTIs. Luo et al. (2017) proposed a network integration pipeline DTINet to integrate heterogeneous data sources (e.g., drugs, proteins, diseases and side-effects). Most of the above methods rely on traditional machine learning, ignoring the topology information of drugs and proteins, and unable to learn the deep interactions between them.

In recent years, with the great success of deep learning in various fields of bioinformatics, network representation learning methods, which can learn rich topological information and the complex interaction between heterogeneous data, have been used in DTI prediction (Sun et al., 2020; Zeng et al., 2020). Peng et al. (2020) proposed a learning based method DTI-CNN for drug-target interaction prediction that learns low-dimensional vector representations of features from heterogeneous networks. DTI-CNN takes the concatenate representation vector of drugs and proteins as input, and adopts convolution neural networks (CNN) as the classification model. Nevertheless, DTI-CNN ignores the interactions between drug and protein pairs (DPPs) in the modeling and learning process. To incorporate associations between DPPs into DTI modeling, Zhao et al. (2020) built a DPP network based on multiple drugs and proteins in which DPPs are the nodes and the associations between DPPs are the edges of the network. They then proposed a model GCN-DTI for DTI identification. The model first uses a graph convolutional network to learn the representation for each DPP and applies a deep neural network to predict the final label based on the representation. Although this method noticed that neighbor DPPs may have an influence on each other, it ignored the semantic feature of the DPP network. At the same time, whether the noise in the graph can be removed, or whether the model can learn more effective representation from limited labeled data, is still unresolved.

Nevertheless, in many real applications such as DTI prediction, labeled data is limited and expensive to obtain. Recently, self-supervised contrastive learning has attracted considerable attention in many graph representation learning tasks. Current works mainly focus on designing different graph augmentation strategies to produce two representations of the same node and leverage a contrastive learning loss to maximize agreement between them, while minimizing the similarity between “other negative samples” (You et al., 2020). GraphCL (Hafidi et al., 2020) learns node embeddings by maximizing agreement/similarity between the representations of two randomly perturbed versions of the same node's local subgraph. You et al. (2020) designed four types of graph data augmentation (node dropping, edge perturbation, attribute masking, and subgraph) and studied the influence of various combinations of data augmentation in different tasks. Zhu et al. (2020) combined many kinds of prior knowledge to enhance interrelationship to unimportant nodes and edges, and retain the internal structure and attribute information of the graph. In the field of bioinformatics, graph contrastive learning is applied to DDI prediction, comparing the original graph representation with the processed graph representation (Wang et al., 2021b). At the same time, Ciortan and Defrance (2021) used two encoders to randomly mask the sequences in the scRNA-seq and conducted contrastive learning on the augmented data. However, little effort has been made to fully leverage valuable label information to supervise the construction of effective positive and negative pairs in the contrastive loss.

Probably most relevant work is the model of Wang et al. (2021a) which uses two views of a heterogeneous information network (network schema and meta-path views) to learn node embeddings. Then the cross-view contrastive learning is proposed to extract the positive and negative embeddings from two views. The key difference is that we try to utilize the valuable supervision information, and seek the commonalities between examples of each class and contrast them with examples from other classes (Khosla et al., 2020; Gunel et al., 2021). Through the contrast of the topology structures and semantic features of the DPP network, the two views collaboratively supervise each other and guide the model optimization in a supervised manner.

As illustrated in Figure 1, motivated by the intuition that the similarity between first-order neighbor nodes in one class is higher, we propose an end-to-end Supervised Graph Co-contrastive Learning model for DTI prediction, namely SGCL-DTI. Specifically, SGCL-DTI first learns low dimensional representations of drugs and proteins from heterogeneous networks through a meta-path guided graph encoder. After that, a topology graph and a semantic graph for drug-protein pair (DPP) nodes are constructed respectively. Different from previous contrastive learning which contrasts the original network and the corrupted network, we optimize the final prediction objective function through the co-contrastive learning of the two DPP networks, by maximising the similarity between the first-order neighbors in one class and contrasting them with nodes in other classes (Gunel et al., 2021). We conduct comprehensive experiments on three public datasets and show that our model outperforms the SOTA methods significantly.

The key contributions of this paper can be summarized as follows:

- We propose a supervised graph co-contrastive learning model for the task of DTI prediction. Through contrastive learning of multiple views while incorporating the supervision information, SGCL-DTI significantly outperforms the SOTA methods over all the datasets. To the best of our knowledge, this is the first attempt that contrasts topology structures and semantic features of the same graph in a supervised manner.
- We thoroughly design and conduct comprehensive experiments to prove the effectiveness of different components in our model, and also analyze different positive and negative sample selection strategies in contrastive learning on this task.
- SGCL-DTI is a generic end-to-end graph representation learning framework and can be easily extended to other applications.

2 Materials and methods

In this section, we present our model for DTI prediction and explain the motivation and the key idea behind our model. The architecture of our proposed model, SGCL-DTI, is shown in Figure 2. It consists of four components: Heterogeneous Information Network Construction, Meta-path based Graph Encoder, DPP Network Representation Learning, and Supervised Co-contrastive Optimization. First, to model different types of entities and their complex relationships, we utilize a heterogeneous information network to depict drugs, proteins, and corresponding heterogeneous relations among them. Next, a meta-path based graph encoder with the attention mechanism is used to learn the representation of drug and protein nodes from the heterogeneous information network. Then, the topology network and the semantic network of DPPs are constructed and learned, where each DPP node is the
concatenation of a drug and a protein representation. In this way, we can learn not only the relationship between DPP nodes, but also the interaction within drug-protein pairs. Finally, a novel contrastive optimization module is used to generate a collaborative contrastive loss of the two views, and guide the model optimization in a supervised manner.

2.1 Heterogeneous Information Network Construction

A Heterogeneous Information Network $G = (\mathcal{V}, \mathcal{E})$ is a graph where $\mathcal{V}$ is a set of nodes and $\mathcal{E}$ is a set of edges. $G$ is associated with a node type mapping function $\phi: \mathcal{V} \rightarrow \mathcal{N}$ and an edge type mapping function $\psi: \mathcal{E} \rightarrow \mathcal{R}$. $\mathcal{N}$ and $\mathcal{R}$ denote sets of object and link types, where $|\mathcal{N} + \mathcal{R}| > 2$.

Inspired by the previous work (Luo et al., 2017), we collect diverse information from the public databases to construct the heterogeneous network for our DPP prediction task, among which there are four drug-related networks (drug-drug relationship network, drug-related disease network, drug and side-effect network, and drug-chemical structure similarity network), three protein related networks (protein-related disease network, protein-protein relationship network, and protein-sequence similarity network) and one drug-protein interaction network used as our ground-truth. In this study, we model the drug and protein data as an HIN $G$. Specifically, the constructed HIN $G$ includes four entities (i.e., drug, protein, disease and side-effect) and a series of relationships among them.

2.2 Meta-path based Graph Encoder

Traditional GCNs can only capture the homogeneous relationships in homogeneous networks, which overlook the rich information among them. To address this issue, meta-paths are used to capture the heterogeneous context information in a heterogeneous graph (Fu et al., 2016; Gong et al., 2020). We learn the representation of drugs and proteins according to the pre-defined meta-paths with a graph encoder. Considering that different meta-paths may have different influences on the final representation learning of drugs or proteins, we utilize the attention mechanism to fuse the representation of drugs or proteins learned under the guidance of different meta-paths to generate the joint attention representation.

2.2.1 Meta-path Selection

A Meta-path in $G$ is defined on a network schema $S = (\mathcal{N}, \mathcal{R})$ and is denoted as a path in the form of $N_1 \rightarrow R_1 \rightarrow N_2 \rightarrow R_2 \rightarrow \cdots \rightarrow R_l \rightarrow N_{l+1}$, which describes a composite relation $R = R_1 \circ R_2 \circ \cdots \circ R_l$ between node types $N_1$ and $N_{l+1}$, $\circ$ denotes the composition operator on relations.

Typical meta-paths between two drugs can be defined as follows: $\text{treat} \rightarrow \text{disease} \rightarrow \text{treat}^{-1} \rightarrow \text{drug}$, which means that two different drugs are related because they can treat the same disease; $\text{cause} \rightarrow \text{side} \rightarrow \text{effect} \rightarrow \text{drug}$, which denotes that two drugs are related in containing chemical components that cause the same side effects. On the other hand, a meta-path between two kinds of proteins can be defined as $\text{protein} \rightarrow \text{express} \rightarrow \text{disease} \rightarrow \text{express}^{-1} \rightarrow \text{protein}$, which denotes that two kinds of protein are related due to the abnormal expression on the same disease. In total, we induce five meta-paths for drugs and four meta-paths for proteins from $S$, as shown in Figure 2.

2.2.2 Meta-path Guided GCN

After the appropriate meta-paths are selected, we learn the representations of drugs and proteins based on different meta-paths through GCN. Given an HIN $G = (\mathcal{V}, \mathcal{E})$ with a group of meta-paths $P = \{P_1, P_2, \ldots, P_M\}$ and the corresponding adjacency matrix $A = \{A_1, A_2, \ldots, A_M\}$ ($M$ represents the number of meta-paths).

First, we use a multi-layer GCN to generate the drug representation based on each meta-path of the drug as follows:

$$z_{p_i}^{(l)} = ReLU \left( \frac{1}{|\mathcal{N}|} \mathbf{A}_p \mathbf{D}_p^{-\frac{1}{2}} z_{p_i}^{(l-1)} \mathbf{W}_p^{(l)} \right)$$

where $ReLU(x) = \max(0, x)$ is a rectified linear activation function. $z_{p_i}^{(l)}$ denotes the $(l)$-Layer representation of drug, in which the subscript $i$ means the $i$-th meta-path of drugs. Specially, $z_{p_i}^{(0)}$ is the input vector which is initialized from a standard normal distribution. We add an identity matrix $I_p$ to the adjacency matrix $A_p$ to indicate the node itself, where $A_p = A_p + I_p$. $\mathbf{D}_p$ is the diagonal matrix of $\mathbf{A}_p$, $\mathbf{W}_p \in \mathbb{R}^{d_{in} \times d_{out}}$ is the shared trainable weight matrix for both drugs and proteins at layer $l$. $d_{in}$ and $d_{out}$ represent the input dimension and output dimension of each layer of GCN respectively.

In order to learn the different influences of different meta-paths for the final representation of drugs, given the corresponding representation of each meta-path of drugs $\mathbf{e}_p \in \{e_{p_1}, e_{p_2}, \ldots, e_{p_M}\}$, we can learn the attention weights and obtain the final drug representation $\mathbf{h}_{\text{drug}}$ as follows:

$$\mathbf{h}_{\text{drug}} = \sum_{i=1}^{M} \alpha_{p_i} \mathbf{e}_p$$

where $\mathbf{W} \in \mathbb{R}^{d_j \times d_p}$ is a weight matrix, $\mathbf{b} \in \mathbb{R}^{d_j \times 1}$ is a bias vector and $\mathbf{q} \in \mathbb{R}^{d_j \times 1}$ is a shared attention vector. $d_j$ is the dimension of the attention layer, $d_p$ is the dimension of $\mathbf{e}_p$. The obtained weight values are normalized by $\text{softmax}$ function as attention scores $\alpha_{p_i} = \text{softmax}(\alpha_{p_i})$.

Similarly, we can obtain the protein representation $\mathbf{h}_{\text{protein}}$ based on each meta-path of the proteins.

2.3 DPP Network Representation Learning

In order to capture the deep and comprehensive relationships between drugs and proteins, we combine each drug $p$ and protein $q$ to form a drug and
First, we build the topology graph $G_t = (A_t, X_{DPP})$, where $X_{DPP} \in \mathbb{R}^{N_{DPP} \times d_{DPP}}$ is the representation matrix of all the DPP nodes, $N_{DPP}$ represents the total number of DPPs, $d_{DPP}$ is the dimension of a DPP representation vector. We follow the principle that if two DPPs contain a common drug or a common protein, there is an edge between them. The adjacency matrix $A_t \in \mathbb{R}^{N_{DPP} \times N_{DPP}}$ represents the relationships of edges between nodes in a graph. The value of element in the $i$-th row and the $j$-th column of $A_t$ equals to 1, meaning that the two DPPs share some common features, and it is 0 in contrary.

2.3.2 Semantic Graph Construction

We construct the semantic graph of DPPs based on the semantic similarity of learned representations. For each DPP $i$, we calculate the cosine similarity of the representation between $i$ and other DPPs, and then select the top $K$ nearest DPP nodes as its adjacent nodes (Kipf and Welling, 2016). For example, if DPP $j$ and DPP $k$ are adjacent nodes of DPP $i$, then we can define in the adjacent matrix $A_s$, the element in the $i$-th row and the $j$-th column and the element in the $i$-th row and the $k$-th column to be 1, and 0 otherwise. Therefore, we can construct a semantic graph of DPPs $G_s = (A_s, X_{DPP})$, where $A_s$ is the adjacency matrix.

2.3.3 DPP Network Learning

We use two multi-layer GCNs to learn the representation of two DPP networks. Let $z_t$ and $z_s$ denote the DPP node representation learned from the topology graph and the semantic graph respectively, the output of the $l$-th layer of GCN models is as follows:

$$z_{t}^{(l)} = ReLU(D_t^{-\frac{1}{2}} \tilde{A}_t D_t^{-\frac{1}{2}} z_{t}^{(l-1)} W_{t}^{(l)})$$

$$z_{s}^{(l)} = ReLU(D_s^{-\frac{1}{2}} \tilde{A}_s D_s^{-\frac{1}{2}} z_{s}^{(l-1)} W_{s}^{(l)})$$

where $\tilde{A}(i,j) = A(i,j)+I(i,j)$ is an identity matrix, $\tilde{D}(i,j)$ is the diagonal degree matrix of $\tilde{A}(i,j)$ and $W_{t}^{(l)}$, $W_{s}^{(l)}$ is the weight matrix of the $l$-th layer of GCN.

2.4 Supervised Co-contrastive Optimization

In this paper, we propose a supervised contrastive learning objective to minimize intra-class variance by pulling together examples belonging to the same class and maximize inter-class variance by pushing apart samples from different classes. The supervised contrastive learning (SCL) strategy in SGCL-DTI is similar to the contrastive objectives used in self-supervised graph contrastive learning. The difference is that the contrastive objective is used to supervise the learning of the final prediction task, instead of contrasting different augmented views of examples.

Hence, the objective in SGCL-DTI includes a supervised classification term and a contrastive learning term for DTI prediction. When calculate the contrastive loss, we need to define positive and negative samples. Different from previous work, we propose a new positive sample selection strategy based on the topology network and the semantic network. Specifically, given a DPP node $i$ in the topology network, we not only take its corresponding node $j$ in the semantic network as a positive sample, but also take node $j$’s first-order neighbor nodes with the same class label as positive samples. We think that first-order neighbor nodes with the same class label are highly correlated. Hence, we can realize the collaborative contrastive learning between the two graphs.

The supervised co-contrastive learning loss of the topology graph $L_{SCL}^{t}$ can be defined as follows:

$$L_{SCL}^{t} = -\sum_{i=1}^{N_{DPP}} \log \frac{\exp \left( \frac{\langle z_{t}^{i}, z_{t}^{i} \rangle}{\tau} \right)}{\sum_{k \in (P_i \cup N_i)} \exp \left( \frac{\langle z_{t}^{i}, z_{k}^{i} \rangle}{\tau} \right)}$$

where $P_i$ is the set of positive samples of DPP node $i$ ($i \in G_t$), which is composed of node $j$ ($j \in G_s$) and node $j$’s first-order neighbor nodes with the same class label. $N_i$ is the set of negative samples of node $i$, representing all the other nodes not in $P_i$. $\tau$ is an adjustable scalar temperature parameter. Similarly, we can obtain the supervised co-contrastive learning loss of the semantic graph $L_{SCL}^{s}$.

We model the task of DTI prediction as a binary classification problem, that is to predict whether there is an interaction between a DPP pair $x_i$. The loss function of classification can be defined as:

$$L_{CE} = -\sum_{i=1}^{N_{DPP}} [y_i \log \pi(x_i) + (1 - y_i) \log (1 - \pi(x_i))]$$

where $\pi(x) = P(Y = 1 | x)$ and $y \in \{0, 1\}$.

Finally, the optimization objective of our model consists of three parts: the classification loss, the contrastive loss and the L2 regularization term $R(\Theta)$:

$$L = L_{CE} + \lambda L_{SCL}^{t} + (1 - \lambda) L_{SCL}^{s} + \gamma R(\Theta)$$

where $R(\Theta) = \sum(\Theta)^2$ (i.e., the sum of the squared weight values) , $\Theta$ represents all the trainable model parameters.

3 Results

In this section, we first introduce the datasets, comparison methods, and evaluation metrics used in the experiment. Then, we conduct experiments to answer the following questions: (1) Is it feasible and effective to predict the DTIs based on the proposed SGCL-DTI? (2) Is it useful to incorporate supervised contrastive learning and co-contrastive learning into the framework? (3) Is our strategy of selecting positive and negative samples effective comparing to others? (4) What is the effect of our method for special cases such as cold-start DPPs?

3.1 Experimental Setup

3.1.1 Dataset

To evaluate the performance of the end-to-end heterogeneous graph representation learning-based framework for DTI prediction, we test our model on three widely used datasets in previous studies, namely, Zheng’s DTIdata, Luo’s DTIdata and Yamanishi’s DTIdata. These three data sets belong to large, medium and small data sets in scale.

- Zheng’s DTIdata (Zheng et al., 2018) There were 11,819 experimentally verified interactions between 1,094 drugs and 1,556 target proteins in the data set. In addition, besides drugs and targets it also contains four kinds of heterogeneous information: drug substitute, chemical structure, drug side-effect and gene ontology.
- Luo’s DTIdata (Luo et al., 2017) There are four types of nodes (drug, protein, disease and side-effect) and eight types of interactions constructing totally six heterogeneous interaction types (drug-drug interaction, drug-protein interaction, drug-disease interaction, drug-side effect interaction, protein-protein interaction and protein-disease interaction) and two similarity networks (pharmacochemical structure similarity network and protein sequences similarity network), which covers 12,015 nodes and 1,894,854 edges in total.
- Yamanishi’s DTIdata (Yamanishi et al., 2010) There are four sub-datasets (nuclear receiver (NR), G-protein-coupled receptor (GPCR), ion channels (IC) and enzyme) including 1,481 drugs and 1,408 proteins, with a total of 9,880 DTIs. Besides, each sub-dataset contains a drug-drug similarity network and a protein-protein similarity network.
For comparison, we use the following competitive methods for DTI prediction as baselines:

- **DTINet** (Luo et al., 2017) A novel network integration pipeline for DTI prediction. DTINet learns low-dimensional feature vectors of drugs and targets across multiple networks, and then finds an optimal projection from the drug space onto the target space and predicts the interactions.

- **GCN-DTI** (Zhao et al., 2020) Graph convolutional network for DTI prediction. In order to incorporate the association within a drug protein pair, GCN-DTI constructs a DPP network based on multi-drugs and proteins, which takes DPP as the node and the association between DPP as the edge of the network. On this basis, the model uses a graph convolution network to learn the characteristics of each DPP and the final label is predicted by using a deep neural network.

- **IMCHGAN** (Li et al., 2021) Inductive Matrix Completion with Heterogeneous Graph Attention Network. IMCHGAN proposes a two-level graph attention network to learn drug and target latent feature representations from the DTI heterogeneous network separately, and adopts the inductive matrix completion to predict the DTIs.

- **EEG-DTI** (Peng et al., 2021) An end-to-end learning framework based on heterogeneous graph convolutional networks for DTI prediction. EEG-DTI constructs a heterogeneous network by combining eight types of biological networks, and using the inner product method to calculate the interaction score between drugs and targets based on the learned low-dimensional representations.

### 3.1.3 Experiment Settings

For all the comparison methods mentioned above, we carry out experiments on the same data set with our model and follow the parameter settings in their papers. For the model, the number of GCN layers is $n$, the dimensions of the two hidden layers are $256$ and $128$, respectively. We use a Adam algorithm to train the model for 1000 iterations with early stop. The learning rate is $1e-4$, with a weight decay rate of $1e-10$, and the dropout rate is 0.5. The two hidden layers are $256$ and $128$, respectively. We use an Adam algorithm to train the model for 1000 iterations with early stop. The learning rate is $1e-4$, with a weight decay rate of $1e-10$, and the dropout rate is 0.5.

### 3.2 Experiment Results

#### 3.2.1 Comparison with Baselines

We compare our model and the baselines on the task of DTI prediction, the AUROC and AUPR results are shown in Table 1. We can draw the following conclusions: (i) The methods based on graph neural networks (GCN-DTI, IMCHGAN, EEG-DTI and SGCL-DTI) generally perform better than the deep learning method DTINet, which shows that graph neural network has advantages in learning the interactions from heterogeneous networks. According to this result, we analyze that the graph neural networks learning to update the central node by aggregating the neighbor information of nodes, so that the model can learn the interactions between nodes and more effective node representations, and finally improving the node classification task. (ii) Either only using the topology graph or the semantic graph, the results decrease by about 8% comparing to the full model. This result shows that multi-view is effective for DTI prediction. (iii) Comparing the results of “Without contrastive learning” and SGCL-DTI, the contrastive learning strategy improves the results on the two data sets by nearly 2%.

#### 3.2.2 Ablation Study

To test the effectiveness of different components in our model, we conduct ablation study on the two largest data sets. Specifically, we denote our method as the full model SGCL-DTI and perform the leave-one-out validation on each part of the model. The model “Without contrastive learning” denotes our model which uses meta-path and multi-view GCN, but without contrastive learning. The model “Without semantic graph” denotes our model using only the topology graph and contrastive learning within one graph. The model “Without contrastive learning” denotes our model that removes both contrast learning and multi-view graph learning.

From Table 2 and Table 3, we can draw the following conclusions: (i) The methods based on graph neural networks (GCN-DTI, IMCHGAN, EEG-DTI and SGCL-DTI) generally perform better than the deep learning method DTINet, which shows that graph neural network has advantages in learning the interactions from heterogeneous networks. According to this result, we analyze that the graph neural networks learning to update the central node by aggregating the neighbor information of nodes, so that the model can learn the interactions between nodes and more effective node representations, and finally improving the node classification task. (ii) Either only using the topology graph or the semantic graph, the results decrease by about 8% comparing to the full model. This result shows that multi-view is effective for DTI prediction. (iii) Comparing the results of “Without contrastive learning” and SGCL-DTI, the contrastive learning strategy improves the results on the two data sets by nearly 2%.
Table 4. Results of cold start analysis in Luo’s DTI data and Zheng’s DTI data, where $N$ represents the number of interactions, and $C$ represents the number of cold start nodes

<table>
<thead>
<tr>
<th>Method</th>
<th>Luo’s DTI data</th>
<th>Zheng’s DTI data</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$N = 0$</td>
<td>$N \leq 3$</td>
</tr>
<tr>
<td>$C = 37$</td>
<td>$C = 503$</td>
<td>$C = 1074$</td>
</tr>
<tr>
<td>GCN-DTI</td>
<td>0.7837</td>
<td>0.7912</td>
</tr>
<tr>
<td>IMCHGAN</td>
<td>0.8618</td>
<td>0.8767</td>
</tr>
<tr>
<td>EEG-DTI</td>
<td>0.8649</td>
<td>0.8767</td>
</tr>
<tr>
<td>SGCL-DTI</td>
<td><strong>0.8919</strong></td>
<td><strong>0.9284</strong></td>
</tr>
</tbody>
</table>

When we remove co-contrastive learning across multiple graphs “Without co-contrastive learning”, the result is the worst.

3.2.3 Selection Strategy of Positive and Negative Samples

In contrastive learning, the selection strategy of positive and negative samples will affect the final result. Therefore, in this part, we will compare several strategies widely used in previous work. Given that we have two DPP networks $G_t$ and $G_s$, and two augmented views $G_t$ and $G_s$ obtained by edge perturbation and attribute masking:

- **Strategy S1**: Using a DPP node $i$’s corresponding node $j$ in the opposite network as the positive sample.
- **Strategy S2**: Using a DPP node $i$’s corresponding node $j$ in its augmented view as the positive sample.
- **Strategy S3**: For a DPP node $i$ in one augmented view, using its corresponding node $j$ in another augmented view as the positive sample.

From Figure 3, we find that the contrastive learning methods between the topology graph and the semantic graph (S1 and our method SGCL-DTI), outperform the contrastive learning between one graph and its augmented view (S2 and S3). Furthermore, our method has achieved the best results due to the introduction of supervision information and co-contrastive learning.

![Fig. 3. Comparison of selection strategy of positive and negative samples.](image)

3.2.4 Cold Start Analysis

Previous experiments can prove that the supervised graph co-contrastive learning can greatly improve the results of DTI prediction. However, when there is little information about the interaction between a drug and a protein node, the learning ability of most models may decrease greatly. This is a very common cold start scenario in real-world applications. Therefore, we conduct the cold start analysis. We select the DPP nodes with less than or equal to 0, 3, 5 relationships in Luo’s DTI data, and the DPP nodes with less than or equal to 3, 5, 10 relationships in Zheng’s DTI data as cold start nodes respectively. The results in Table 4 show that our model still outperforms other baseline methods under cold start conditions, and, interestingly, the improvement is even more obvious.

![Fig. 4. Results of parameter sensitivity analysis with different $\lambda$.](image)

3.2.5 Parameter Sensitivity Analysis

Next, we conduct the parameter sensitivity experiment of our model. First of all, we vary the value of the coefficient $\lambda$ in the loss function (Equation 8), to explore the influence of the contrastive learning loss on the model. From Figure 4, we try the following values of $\lambda = \{ 0, 0.2, 0.4, 0.6, 0.8, 1 \}$, and find that the performance improves when $\lambda$ increases. The best results are achieved when $\lambda$ is close to 0.8.

Although there are a large number of drugs and targets, the interactions between drugs and targets has rarely been verified. Aiming at this imbalance of data, we design experiments under different negative case ratios. We tested the positive and negative ratios of 1:1, 1:5 and 1:10 respectively to verify the generalization ability of the model when the number of negative cases is much greater than that of positive cases. From the results in Table 6, we can see that even if the positive and negative examples are unbalanced, the results of AUROC and AUPR of our model are still very stable.

3.2.6 Robustness Test

Finally, we conduct robustness test of our model. We try to find the positive DTI samples $\Delta$ with high homology following the strategies in Luo’s paper: (1) DTIs involving homologous proteins (sequence recognition score
Table 7. Example of robustness test result in Luo’s DTI data

<table>
<thead>
<tr>
<th>Proportion</th>
<th>AUROC</th>
<th>AUPR</th>
<th>AUROC</th>
<th>AUPR</th>
<th>AUROC</th>
<th>AUPR</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>fold 1</td>
<td>0.9807</td>
<td>0.9787</td>
<td>0.9767</td>
<td>0.9560</td>
<td>0.9838</td>
<td>0.9456</td>
</tr>
<tr>
<td>fold 2</td>
<td>0.9664</td>
<td>0.9754</td>
<td>0.9634</td>
<td>0.9367</td>
<td>0.9716</td>
<td>0.9340</td>
</tr>
<tr>
<td>fold 3</td>
<td>0.9727</td>
<td>0.9758</td>
<td>0.9807</td>
<td>0.9531</td>
<td>0.9852</td>
<td>0.9432</td>
</tr>
<tr>
<td>fold 4</td>
<td>0.9804</td>
<td>0.9793</td>
<td>0.9843</td>
<td>0.9602</td>
<td>0.9897</td>
<td>0.9568</td>
</tr>
<tr>
<td>fold 5</td>
<td>0.9857</td>
<td>0.9746</td>
<td>0.9903</td>
<td>0.9690</td>
<td>0.9943</td>
<td>0.9611</td>
</tr>
<tr>
<td>Average</td>
<td>0.9771</td>
<td>0.9768</td>
<td>0.9791</td>
<td>0.9550</td>
<td>0.9849</td>
<td>0.9481</td>
</tr>
</tbody>
</table>

> 40%); (2) DTIs with similar drugs (Tanimoto coefficient > 60%); (3) DTIs of which the drug had similar side effects (Jaccard similarity score > 60%); (4) DTIs with drugs or proteins related to similar diseases (Jaccard similarity score > 60%); (5) DTIs with similar drugs (Tanimoto coefficient > 60%) or homologous proteins (sequence consistency score > 40%).

Then we conduct the following experiments on $\Delta$ : (1) Remove the positive DTI samples with high homology from the training set, and see if the model can predict these DTI interactions in the test phase; (2) Use the positive DTI examples with high homology as negative examples in the training set, to see if the model can predict DTI interactions in the testing stage. In this way, we can simulate the false negative examples of DTIs. For the first strategy, we successfully predicted 673 interactions out of 685 removed DTIs, with an accuracy of 98.2%, while the second strategy successfully predicted 606 interactions, with an accuracy of 98.5%. Through the above experiments, we find that our model is robust in the case of false negatives, and can still predict DTI interactions with high homology. We randomly selected two drugs and their related targets were predicted correctly, and showed the results in Table 7.

Table 7. Example of robustness test result in Luo’s DTI data

<table>
<thead>
<tr>
<th>Drug-Proteins</th>
<th>Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>Olanzapine - CHRM3</td>
<td>True</td>
</tr>
<tr>
<td>Olanzapine - DRD5</td>
<td>True</td>
</tr>
<tr>
<td>Olanzapine - CHRM5</td>
<td>True</td>
</tr>
<tr>
<td>Olanzapine - ADRA2C</td>
<td>True</td>
</tr>
<tr>
<td>Olanzapine - CHRM4</td>
<td>True</td>
</tr>
<tr>
<td>Olanzapine - HTR2C</td>
<td>True</td>
</tr>
<tr>
<td>Olanzapine - ADRA2B</td>
<td>True</td>
</tr>
<tr>
<td>Olanzapine - ADRA1B</td>
<td>True</td>
</tr>
<tr>
<td>Olanzapine - DRD3</td>
<td>True</td>
</tr>
<tr>
<td>Olanzapine - HTR1D</td>
<td>True</td>
</tr>
</tbody>
</table>

Accuracy 100%

<table>
<thead>
<tr>
<th>Drug-Proteins</th>
<th>Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pergolide - DRD5</td>
<td>True</td>
</tr>
<tr>
<td>Pergolide - ADRA2C</td>
<td>True</td>
</tr>
<tr>
<td>Pergolide - HTR2C</td>
<td>True</td>
</tr>
<tr>
<td>Pergolide - ADRA2B</td>
<td>True</td>
</tr>
<tr>
<td>Pergolide - ADRA1B</td>
<td>True</td>
</tr>
<tr>
<td>Pergolide - DRD3</td>
<td>True</td>
</tr>
<tr>
<td>Pergolide - HTR1D</td>
<td>True</td>
</tr>
<tr>
<td>Pergolide - ADRA1D</td>
<td>True</td>
</tr>
</tbody>
</table>

Accuracy 100%

4 Discussion

In this paper, we propose an end-to-end supervised graph co-contrastive learning model for the task of DTI prediction. Our model contrasts topology structures and semantic features of the DPP network and benefits from the valuable supervision information. We conduct comprehensive experiments and demonstrate significant improvements over competitive baselines on three public datasets. We also show that our proposed objective makes the model more robust, especially in the cold start scenarios.

Data availability

Experimental data sets and experimental codes can be found in https://github.com/catly/SGCL-DTI.

Conflict of Interest

We declare that we do not have any conflicts of interest in this paper.

Acknowledgements

We thank the anonymous reviewers for their constructive suggestions. This work was supported by the National Natural Science Foundation of China [62072095, 61806049, 61771165] and the Heilongjiang Postdoctoral Science Foundation [LBH-Z201014]. Corresponding author: Guohua Wang, E-mail: ghwang@nefu.edu.cn.

References


