Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: Day Month Year
Manuscript Category

OXFORD

Subject Section

HGNNLDA: Predicting lncRNA-drug sensitivity associations via a dual channel hypergraph neural network

Dayun Liu [1], Xiaowen Hu [2], Jiaxuan Zhang [3], Zhirong Liu [4], Lei Deng [1,*]

[1]School of Computer Science and Engineering, Central South University, Changsha, 410083, China
[2]School of Computer Science and Technology, Harbin University of Science and Technology, Harbin, 150080, China.
[3]Department of Electrical Engineering , University of California San Diego, La Jolla, 92093, United States.
[4]School of Software, Xinjiang University, Urumqi, 830049, China.

*E-mail: leideng@csu.edu.cn.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Drug sensitivity is critical for enabling personalized treatment. Many studies have shown that long non-coding RNAs (lncRNAs) are closely related to drug sensitivity because lncRNAs can regulate genes related to drug sensitivity to affect drug efficacy. Exploring lncRNA-drug sensitivity associations has important implications for drug development and disease treatment. However, identifying lncRNA-drug sensitivity associations based on traditional biological approaches is small-scale and time-consuming. In this work, we develop a dual-channel hypergraph neural network-based method named HGNNLDA to infer unknown lncRNA-drug sensitivity associations. To our best knowledge, HGNNLDA is the first computational framework to predict lncRNA-drug sensitivity associations. HGNNLDA applies the hypergraph neural network to obtain high-order neighbor information on the lncRNA hypergraph and the drug hypergraph, respectively, and utilizes a joint update mechanism to generate lncRNA embeddings and drug embeddings. Unlike traditional graphs that only have connections between two nodes, hypergraphs can well describe the higher-order connectivity of the lncRNA and drug graphs. The comprehensive experimental results show that HGNNLDA significantly outperforms the other six state-of-the-art models. Case studies on two drugs further illustrate that HGNNLDA is an effective tool to predict lncRNA-drug sensitivity associations. Source codes and data are available at: https://github.com/dayunliu/HGNNLDA

1 Introduction

Long non-coding RNAs (lncRNAs) are RNA molecules over 200 nt in length. lncRNAs play critical roles in many biological processes such as epigenetic regulation, cell cycle regulation, cell differentiation, transcriptional and post-transcriptional regulation, and genome splicing Ponting *et al.* (2009); Rinn and Chang (2012); Geisler and Coller (2013). Relevant studies have shown that lncRNAs regulate human diseases Harries (2012) through the joint action of a series of biomolecules in organisms. Their mutations and dysfunctions are closely related to human diseases such as nervous system diseases, blood diseases, cardiovascular diseases, and various cancers Mercer and Mattick (2013); Gupta *et al.* (2010). With sequencing technology development, more and more lncRNA molecules have been detected and analyzed in sensitivity and depth Yang *et al.* (2010), especially their role in drug sensitivity Bhat *et al.* (2020). Studies have shown that lncRNAs can modulate drug sensitivity-related genes, induce alternative signaling pathways and further affect drug efficacy Hahne and Valeri (2018). For example, lncRNA NORAD inhibits the proliferation of osteosarcoma HOS/DDP cells and increases their sensitivity to cisplatin by targeting miR-410-3p Xie *et al.* (2020). Gallbladder cancer chemotherapy induces gallbladder cancer cell sensitivity through key regulator lncRNA1 (GBCDRlnc1)Cai *et al.* (2019). Identifying lncRNA-drug sensitivity associations has important implications for drug development. However, traditional methods based on biological experiments are often small-scale and time-consuming. Therefore, it is urgent to develop computational methods to identify lncRNA-drug sensitivity associations on a large scale.

LncRNA-drug associations prediction is a breakthrough work. No computational method has been proposed to solve this problem. However, there are some association prediction tasks in the field of bioinformatics that we can learn from, such as lncRNA-disease association prediction, miRNA-disease association prediction, circRNA-disease association prediction, drug-disease association prediction, and microbe-disease association prediction. Based on lncRNAs with similar functions that are more related to diseases with similar phenotypes Sharan and Ideker (2008); Lu *et al.* (2008), Sun et al.Sun *et al.* (2014) proposed a method named RWRlncD to predict potential lncRNA-disease associations. RWRlncD constructed a lncRNA functional similarity network, and employed the random walk with restart on the lncRNA functional similarity network to infer potential lncRNA-disease associations. Li et al.Li *et al.* (2018) proposed a label propagation model named LPLNS based on linear neighborhood similarity to predict unknown miRNA-disease associations. LPLNS computed pairwise linear neighborhood similarities between miRNAs and pairwise linear neighborhood similarities between diseases, respectively, along with the known miRNA-disease associations, are fed into a label propagation method to score each disease-miRNA pair. Li et al.Li *et al.* (2020) use DeepWalk, a network embedding method, on the known circRNA-disease association network to learn the embeddings of nodes to predict circRNA-disease association. Compared to similar methods that predicted circRNA-disease association, Li's method had more flexibility using DeepWalk. Wu et al.Wu *et al.* (2021) proposed an approach named GAERF based on graph auto-encoder and random forest to predict unknown lncRNA-disease association. GAERF used a graph auto-encoder algorithm to learn the representation vectors of nodes from the heterogeneous network with lncRNA, miRNA, and disease and fed them into a random forest classifier. And then, the trained random forest classifier will be applied to predict the lncRNA-disease associations. To capture known complex miRNA and disease data, Xuan et al.Xuan *et al.* (2019) proposed a prediction method called CNNMDA based on network representation learning and convolutional neural networks to predict disease-related miRNAs. CNNMDA via dual-channel to learn the original and global representation of a miRNA-disease pair and the low-dimensional representation of each node on the miRNA-disease association network respectively in the embedding layer. And then, these representations learned in the embedding will be fed to the convolutional modules to deeply learn the complex nonlinear relationship between miRNA and disease. Liu et al. Liu *et al.* (2021) proposed a computational framework for miRNA-disease associations named SMALF. SMALF first extracts latent features from the original features in the miRNA-disease association matrix. Then, combining similar features and latent features is fed into XGBoost to infer unknown miRNA-disease associations. Wang et al.Wang *et al.* (2020) presented a method to uncover disease-related circRNAs based on fast learning with a graph convolutional network algorithm. In Wang's approach, the disease semantic similarity information and known circRNA-disease associations will be fed into the fast learning with graph convolutional network algorithm to extract the high-level features and the forest by penalizing attributes classifier will be used to accurately predict the new circRNA-disease associations. Peng et al.Peng *et al.* (2018) proposed a computational approach based on adaptive boosting to predict microbe-disease associations by scoring the disease-microbe pair using a solid classifier that consists of multiple weak classifiers according to the corresponding weights. Liu et al. Dayun *et al.* (2021) proposed a multi-component graph attention network framework, MGATMDA, which first used a decomposer with an attention mechanism to extract the latent features of the microbiome-disease bipartite graph. Then, these latent features are recombined into a unified embedding. Finally, a fully connected network is used to predict unknown microbial disease associations.

In this work, we develop a computational framework based on hypergraph neural network to predict lncRNA-drug sensitivity associations, called HGNNLDA. First, the known lncRNA-drug sensitivity associations are modeled as a lncRNA-drug bipartite graph. Then, HGNNLDA used the lncRNA-drug bipartite graph to construct lncRNA hypergraph and drug hypergraph, respectively. Subsequently, HGNNLDA combined lncRNA hypergraph and drug hypergraph and used the hypergraph neural network to generate lncRNA and drug embedding. Finally, HGNNLDA uses the inner product to infer the lncRNA-drug sensitivity association. HGNNLDA better explores the higher-order connectivity in the lncRNA-drug bipartite graph than the graph neural network methods. The experimental results show that HGNNLDA takes the highest AUC value compared to the other six models. The case study on two common drugs further validated the effectiveness of HGNNLDA in inferring the unknown lncRNA-drug sensitivity associations.

2 Methods

2.1 Dataset

We obtained lncRNA-drug sensitivity associations from the RNAactDrug Dong *et al.* (2020) database. RNAactDrug is a comprehensive database that provides drug sensitivity associated RNA molecules including lncRNA, miRNA, mRNA from multi-omics data. RNAactDrug has 19,770 mRNAs, 11,119 lncRNAs, 438 miRNAs and 4,155 drugs. After removing redundant information, we constructed a benchmark dataset with 36,248 lncRNA-drug sensitivity associations, including 978 lncRNAs and 1,815 drugs.

2.2 lncRNA-drug bipartite graph

Known lncRNA-drug sensitivity associations can be modeled as a lncRNA-drug bipartite graph. Let $B \in \mathcal{R}^{|L|\times|D|}$ represent the lncRNA-drug sensitivity association matrix, where L and D are the lncRNA set and the drug set, respectively. We can build a lncRNA-drug bipartite graph $G = (L, D, E)$. For any edge $e = (l, d) \in E$, it means that there is an experimentally verified association between drug l and disease d. We use the high-order connectivity in the lncRNA-drug bipartite graph to generate lncRNA hypergraph and drug hypergraph, and use the hypergraph neural network to generate lncRNA embedding and drug embedding.

2.3 HGNNLDA

In this work, we propose a computational method to predict lncRNA-drug sensitivity associations, named HGNNLDA. HGNNLDA takes the lncRNA-drug bipartite graph as input and outputs the probability that lncRNA l is associated with drug d. As shown in Figure 1, HGNNLDA first uses the lncRNA-drug bipartite graph to generate lncRNA hypergraph and drug hypergraph based on defined rules which is a new perspective to understand the original data. Then, HGNNLDA uses hypergraph convolution to learn the complex correlation information of higher-order neighbors from the lncRNA hypergraph and the drug hypergraph, respectively. Subsequently, HGNNLDA designed a joint update mechanism to combine the learned lncRNA high-order neighbor information and drug high-order neighbor information to generate lncRNA and drug embedding. Finally, the inner product was used to infer the association between lncRNA and drug sensitivity. Next, we will introduce HGNNLDA in detail.

2.3.1 Hypergraph construction

In a normal graph, an edge can only be connected to two vertices. An edge is called a hyperedge in a hypergraph, and a hyperedge can be connected to any number of vertices. Like the two-tuple representation of ordinary graphs, a hypergraph is usually defined as $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, $\mathcal{V}$ is the finite

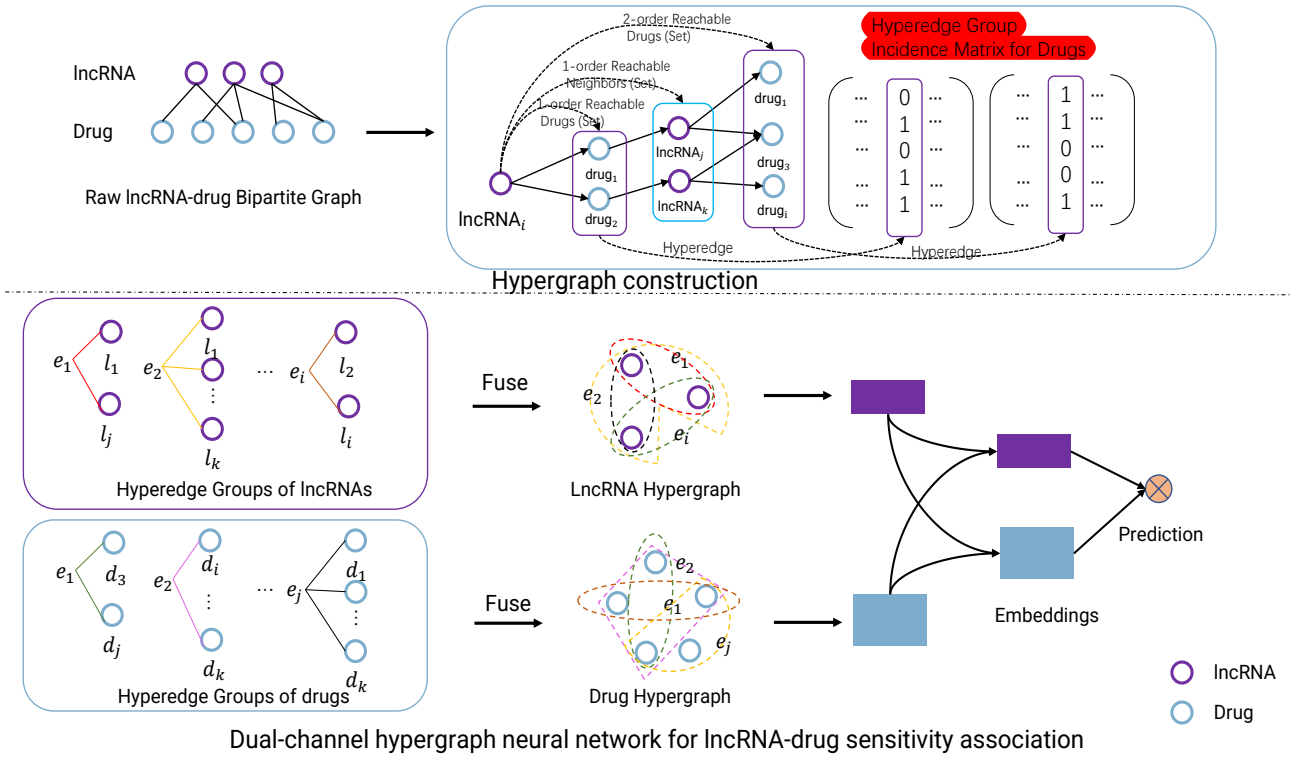


Fig. 1. Overview of HGNNLDA. HGNNLDA first used the higher-order connectivity in the lncRNA-drug bipartite graph to build a lncRNA hypergraph and a drug hypergraph. Subsequently, HGNNLDA uses a hypergraph neural network to learn high-order neighbor information in the lncRNA hypergraph and drug hypergraph. Then, HGNNLDA designs a joint update mechanism to generate lncRNA embeddings and drug embeddings. Finally, HGNNLDA used inner products to predict lncRNA-drug sensitivity associations.

vertices of the hypergraph, and $\mathcal{E}$ is the set of hyperedges the hypergraph. For any hyperedge $e \in \mathcal{E}$, it is a subset of the vertex set $\mathcal{V}$.

Hypergraph $\mathcal{G}$ can be represented by an incidence matrix H of $|\mathcal{V}|\times|\mathcal{E}|$. The rows represent different vertices, and the columns represent different hyperedges. In the incidence matrix H, use $h(\nu, e)$ to indicate whether the vertex ν is on the hyperedge e:

$$h(\nu, e) = \begin{cases} 0 \ \ if \nu \in e \\ 1 \ \ if \nu \notin e \end{cases} \tag{1}$$

In the hypergraph, for each vertex $\nu \in \mathcal{V}$, the degree d(ν) of ν is defined as the number of hyperedges that contain the nodes, expressed as $d(\nu) = \sum_{e\in\mathcal{E}} h(\nu, e)$, for each hyperedge $e \in \mathcal{E}$, $\delta(e)$ of the degree of e is defined as the number of vertices contained on the hyperedge, expressed as $\delta(e) = \sum_{\nu\in\mathcal{V}} h(\nu, e)$. Respectively use the diagonal node degree matrix $D_\nu \in R^{|\mathcal{V}|\times|\mathcal{V}|}$ and the diagonal edge degree matrix $D_e \in R^{|\mathcal{E}|\times|\mathcal{E}|}$ to represent the degree matrix of the vertices and hyperedges, among which the elements on the diagonal. It is the degree of each vertex/hyperedge, namely d(ν) or $\delta(e)$.

Constructing high-order connectivity for lncRNA and drugs respectively to realize the construction of hyperedges on the lncRNA-drug bipartite graph. High-level correlation extraction is performed on lncRNAs and drugs according to custom rules.

On lncRNA. Definition 1: Drug's k-order reachable neighbors. In the lncRNA-drug interaction graph, more specifically a bipartite graph, if there is a series of adjacent vertices (i.e., paths) between $drug_i$ and $drug_j$, then $drug_i$ ($drug_j$) is the k-order reachable neighbor of $drug_j$ ($drug_i$), And the number of lncRNAs of this path is less than k.

Definition 2: Drug's k-order reachable lncRNAs. In the lncRNA-drug bipartite graph, if there is a direct interaction between $lncRNA_j$ and $drug_k$, then $lncRNA_j$ is the k-order reachable neighbor of $drug_i$, and $drug_k$ is the k-order reachable neighbor of $drug_i$.

For drug i, its k-order reachable lncRNAs set is called $B_l^k(d)$. Mathematically speaking, a hypergraph can be defined on a set family, where each set represents a hyperedge. Therefore, the hyperedge here can be constructed from the k-order reachable lncRNA set of the drug. Then we can construct a high-order hyperedge group according to the k-order reachable rule between lncRNAs, which can be expressed as:

$$\mathcal{E}_{B_l^k} = \left\{ B_l^k(d) | d \in D \right\} \tag{2}$$

The k-order reachable matrix of drugs can be expressed as $A_d^k \in \{0,1\}^{M\times M}$, its form is:

$$A_d^k = min\left(1, power\left(H^T \cdot H, k\right)\right), \tag{3}$$

Where power(M, k) is a function for calculating the k power of a given matrix M. $H \in \{0,1\}^{N\times M}$ represents the incidence matrix of the bipartite graph of lncRNA-drug. Then the hyperedge group incidence matrix $H_{B_l^k} \in \{0,1\}^{N\times M}$ constructed by the k-order reachable rule between lncRNAs can be expressed as:

$$H_{B_l^k} = H \cdot A_d^{k-1}. \tag{4}$$

On drugs. The k-order reachable neighbor of lncRNA and the k-order reachable drug of lncRNA can be symmetrically defined similarly. Specifically, the k-order reachable matrix of lncRNA can be expressed as $A_l^k \in \{0,1\}^{N\times N}$, which can be written as:

$$A_l^k = min\left(1, power\left(H \cdot H^T, k\right)\right), \tag{5}$$

The hyperedge group incidence matrix $H_{B_d^k} \in \{0,1\}^{M\times N}$ constructed by the k-order reachable rule between drugs can be expressed as:

$$H_{B_d^k} = H^T \cdot A_l^{k-1}. \tag{6}$$

Hypergraph H_l of lncRNA. Assuming that we have a hyperedge group established on lncRNA by the k-order reachable rule, the final mixed high-order connection between lncRNAs can be represented by the hypergraph $\mathcal{G}_l$ fused with a hyperedge group. Due to the advantages of a hypergraph in multi-modal fusion, the simple concatenation operation of the incidence matrix of hyperedge groups·‖· can be applied to the hyperedge group fusion f(·). Finally, the hypergraph association matrix H_l of lncRNA can be expressed as:

$$H_l = f\left(\mathcal{E}_{B_l^{k_1}}, \mathcal{E}_{B_l^{k_2}}, \cdots, \mathcal{E}_{B_l^{k_a}}\right) = \underbrace{H_{B_l^{k_1}} \| H_{B_l^{k_2}} \| \cdots \| H_{B_l^{k_a}}}_{a} \tag{7}$$

Hypergraph H_d of drug. In the same way, through the k-order reachable rule, b hyperedge groups are constructed. The final hypergraph needs to merge these b hyperedge groups. Finally, the drug's hypergraph incidence matrix Hd can be expressed as:

$$H_d = f\left(\mathcal{E}_{B_d^{k_1}}, \mathcal{E}_{B_d^{k_2}}, \cdots, \mathcal{E}_{B_d^{k_b}}\right) = \underbrace{H_{B_d^{k_1}} \| H_{B_d^{k_2}} \| \cdots \| H_{B_d^{k_b}}}_{b} \tag{8}$$

2.3.2 Hypergraph Convolution

This section will use hypergraph convolution to learn high-order neighboring nodes' complex information from hypergraphs of lncRNA and drug, and generate the embeddings of lncRNA and drug to predict the association.

First, we need to initialize some parameters. We construct ID embedding matrices of lncRNA and drug, and project them to a vector representation. The formula of embedding lookup table formula is shown as follows:

$$\begin{cases} E_l = [e_{l_1}, e_{l_2}, ..., e_{l_n}] \\ E_d = [e_{d_1}, e_{d_2}, ..., e_{d_m}] \end{cases} \tag{9}$$

where n and m are the number of lncRNA and drug, respectively. $e_{l_i} \in R^T$ represents the embedding vector of the ith lncRNA, $e_{d_j} \in R^T$ represents the embedding vector of the jth drug and T represents the embedding size.

And then, we need to aggregate the neighboring message upon pre-construct hypergraph incidence matrix and embedding lookup table. First, we need to normalize the hypergraph incidence matrix H_l, H_d, H_l^T and H_d^T ,they are defined as:

$$\begin{cases} \tilde{H}_l = D_{l_v}^{-1/2} H_l D_{l_e}^{-1/2} \\ \tilde{H}_d = D_{d_v}^{-1/2} H_d D_{d_e}^{-1/2} \end{cases} \tag{10}$$

$$\begin{cases} \tilde{H_l^T} = D_{l_e}^{-1/2} H_l D_{l_v}^{-1/2} \\ \tilde{H_d^T} = D_{d_e}^{-1/2} H_d D_{d_v}^{-1/2} \end{cases} \tag{11}$$

where $D_{l_v}, D_{l_e}, D_{d_v}, D_{d_e}$ are the vertex degrees and hyperedge degrees of lncRNA and drug respectively as mentioned in section 2.3.1. They have no impact on the path of message passing on the hypergraph.

In traditional hypergraph nerual network(HGNN) Feng *et al.* (2018), the hypergraph convolutional layer operation(HGNNConv(.,.)) can be abstracted as:

$$X^{(l)} = \tilde{H}\tilde{H^T}X^{(l)} \tag{12}$$

where $\tilde{H}$ and $X^{(l)}$ is the input of the HGNNConv(.,.), they represent the normalized incidence matrix defined by us and the vertex feature at layer l respectively. $\Theta^{(l)}$ is a trainable parameter. But this method only considers aggregated neighboring node's messages and ignores its original information. Inspired by other methodsHamilton *et al.* (2017); Kipf and Welling (2016); Shervashidze *et al.* (2011), to propagate lncRNA/drug embedding on the hypergraph and retaining original information, we can join the original features to this formula to avoid the problem as follows:

$$X^{(l)} = \tilde{H}\tilde{H^T}X^{(l)} + X^{(l)} \tag{13}$$

And then, the embeddings and hypergraph incidence matrix of lncRNA and drug will be fed into the hypergraph convolutional layer via a dual-channel as follows:

$$\begin{cases} M_l^{(l)} = HGNNConv(\tilde{H}_l, E_l) \\ M_d^{(l)} = HGNNConv(\tilde{H}_d, E_d) \end{cases} \tag{14}$$

After that, the output $M_l^{(l)}$ and $M_d^{(l)}$ would learn the complex correlations from its high-order neighbors respectively.

And then, for distilling discriminative information of lncRNA and drug, we need to jointly update E_l and E_d using $M_l^{(l)}$ and $M_d^{(l)}$ as follows:

$$\begin{cases} E_l^{(l+1)} = \sigma(M_l^{(l)}\Theta^{(l)}) \\ E_d^{(l+1)} = \sigma(M_d^{(l)}\Theta^{(l)}) \end{cases} \tag{15}$$

where $\Theta^{(l)} \in R^{C^l \times C^{l+1}}$ is a trainable parameter and the $C^l \times C^{l+1}$ is the input/output's feature dimension at layer l.

After that, the obtained embeddings of lncRNA and drug will be used to predict the association.

2.3.3 Prediction and Optimization

After getting the embeddings of lncRNA and drug, we will use them to compute specific association scores between lncRNA and drugs, and construct the loss function to optimize the training process.

In prediction part, given a lncRNA l and a target drug d, we can get the association of the lncRNA l to the drug d by computing their inner product:

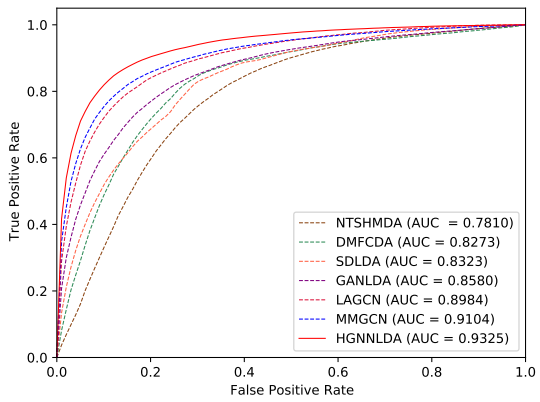
$$\hat{r_{ld}} = e_l^T e_d \tag{16}$$

where e_l and e_d respectively are the embedding obtained from hypergraph neural network, $\hat{r_{ld}}$ is the score of the lncRNA and the target drug.

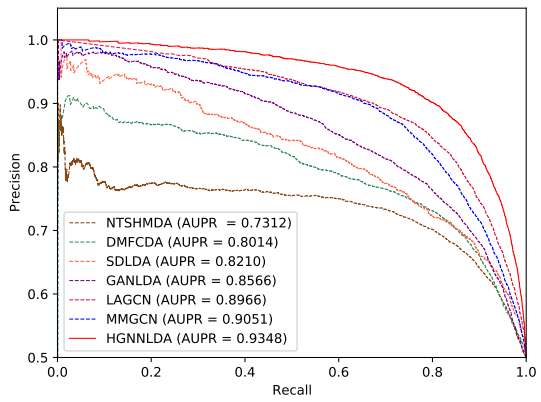
And we will use the Bayesian Personalized Ranking(BPR) for optimization. In our method, only the associations between lncRNA and drugs defined by us will be used to train and predict, and some potential associations remain unknown. So maybe some lncRNA is associated with the target drug, but we cannot get it. Considering that we aim to get the top n drug that is most associated with each lncRNA. Sum consider most rank oriented recommendation, we use BPRRendle *et al.* (2012) for optimization. The loss function is as follows:

$$\ell = -\sum_{(l,d^+,d^-)\in\tau} -ln\sigma(\hat{r_{ld^+}} - \hat{r_{ld^-}}) + \lambda||\Theta||_2^2 \tag{17}$$

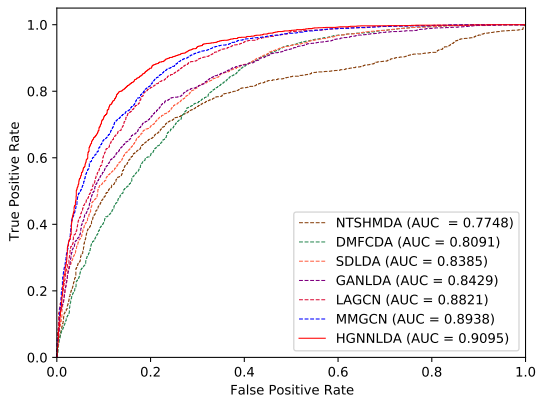
where τ is the trainset, d^+ represents a known association with the lncRNA l, and the association between d^- and the lncRNA l is unknown. $\sigma(.)$ is the logistic sigmoid function. Θ represent the trainable parameters. To avoid overfitting, we introduced the regularization parameter λ.



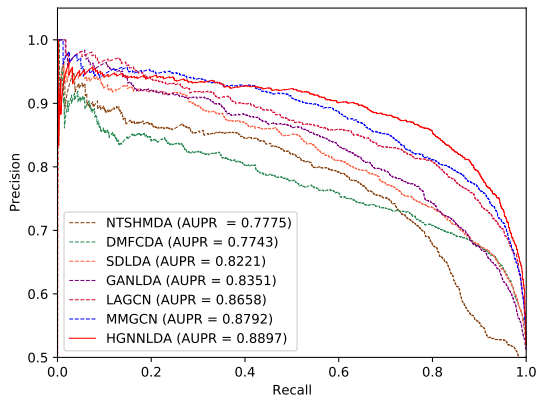
(a) The ROC curves on the benchmark dataset



(b) The PR curves on the benchmark dataset



(c) The ROC curves on the independent dataset



(d) The PR curves on independent dataset

Fig. 2. The ROC and PR curves of all methods on the benchmark dataset and the independent dataset

3 Experiments and results

3.1 Experimental setup

In order to evaluate the performance of HGNNLDA, we performed a five-fold cross-validation experiment on the benchmark dataset. All lncRNA-drug sensitivity associations are randomly divided into five subsets. Each time we select a subset as the test set, and the remaining four subsets as the training set. This process stops until the five subsets are used as the test set. We choose two well-known metrics to evaluate the performance of HGNNLDA including AUC (Liu *et al.* (2020)) and AUPR (Tang *et al.* (2021a)).

3.2 Comparison with highly related methods

Few computational methods have been proposed to predict lncRNA-drug sensitivity associations. Therefore, we compare HGNNLDA with six highly related models which are used in the field of bioinformatics to solve lncRNA-disease associations (SDLDA Zeng *et al.* (2020), GANLDA Wei *et al.* (2021)), miRNA-disease associations (MMGCN Tang *et al.* (2021b)), circRNA-disease associations (DMFCDA Lu *et al.* (2021)), drug-disease associations (LAGCN Yu *et al.* (2021)), microbe-disease association (NTSHMDA Luo and Long (2020)). SDLDA first uses the singular value decomposition algorithm to obtain the linear feature of lncRNA and disease. To further improve the model's performance, SDLDA uses a deep neural network to extract the nonlinear feature of lncRNA and disease from the original lncRNA-disease association matrix and combines linear and nonlinear features to infer unknown lncRNA-disease associations. GANLDA first constructed the lncRNA-disease bipartite graph. Then, GANSDA used principal component analysis to project the original features of lncRNA and disease to the same dimension. Subsequently, the graph attention network was used to extract potential features of lncRNA and diseases. Finally, GANSDA uses a multilayer perceptron to predict the lncRNA-disease association. MMGCN uses graph convolutional neural networks to obtain the features of lncRNA and diseases from multiple perspectives and uses the attention mechanism to fuse these features to generate the final representation of lncRNA and diseases. DMFCDA models circRNA-disease association prediction as a recommendation problem and uses the DMF Xue *et al.* (2017) model in the recommendation system to predict unknown circRNA and disease associations. LAGCN constructed a drug-disease heterogeneous network by integrating multiple similarities and drug-disease associations. LAGCN applies graph convolutional network to drug-disease heterogeneous network to learn the representation of drugs and diseases. NTSHMDA constructed a microbe-disease heterogeneous network and used random walks with restart to predict microbe-disease associations.

We performed five-fold cross-validation to compare these models. The experimental results are shown in Figure 2a and 2b. The AUC

of HGNNLDA reached 0.9325, and the AUPR reached 0.9328, which was higher than the other six excellent models. HGNNLDA showed satisfactory prediction results.

To further evaluate the performance of HGNNLDA, we established an independent test set and compared HGNNLDA with several other excellent models on the independent test set. Specifically, we conducted a literature search in the PubMed database and established an independent test set with 2,093 lncRNA-drug sensitivity associations, including 273 lncRNAs and 480 drugs. We train all models with 36,248 lncRNA-drug sensitivity associations on the benchmark dataset and compare them on the independent test set. Figure 3 summarizes the experimental results of all models on the independent test set. From Figure 2c and 2d, we can see that HGNNLDA achieves 0.9095 AUC, 0.8897 AUPR, which is significantly higher than other models. This can be attributed to HGNNLDA's better modeling of the higher-order connectivity in the lncRNA-drug sensitivity association. The modeling of higher-order associations is crucial in predicting lncRNA-drug sensitivity associations. Only paired associations can be displayed in a normal graph, while in a hypergraph, lncRNA or drugs with high-order associations can be displayed in a hyperedge, and the hypergraph neural network can better handle high-order connectivity. Therefore, HGNNLDA shows a better prediction effect.

3.3 Parameter sensitivity analysis

In HGNNLDA, there are some important parameters that will affect the performance of the model, such as embedding size, the number of layers of the hypergraph neural network. In this part, we look for the most suitable parameters under the five-fold cross-validation experiment.

3.3.1 Effect of Embedding size

Embedding size is an important parameter that will affect the performance of HGNNLDA. Embedding size is a parameter that needs to be manually defined, which represents the dimensions of learned lncRNA and drug features. If the embedding size is too small, HGNNLDA cannot learn the complex relationship between lncRNA and drug sensitivity. If the embedding size is too large, the risk of HGNNLDA overfitting will increase. We evaluate the performance of HGNNLDA by varying embedding in the range of 16, 32, 64, 128, and 256. As shown in Figure 3a, when the embedding size is set to 64, HGNNLDA achieves the best performance.

3.3.2 Effect of Layer

More layers can theoretically learn more complex models, but the risk of over-smoothing will also increase. As the number of layers increases, dissimilar nodes also have similar representations, and it becomes difficult to distinguish between nodes, affecting the model's performance. To study the effect of the number of layers, we changed the number of layers from 1 to 4. As shown in Figure 3b, we can observe that HGNNLDA obtains the best performance when the number of layers is 3.

3.4 Case study

To further verify the effectiveness of HGNNLDA, we selected two common drugs for case studies. Specifically, we first train HGNNLD with the known lncRNA-drug sensitivity association. Then, for two common drugs, we use the trained HGNNLDA to predict candidate lncRNAs. Subsequently, we ranked the candidate lncRNAs in descending order according to the prediction score. Finally, we took out the top 15 lncRNA candidates and searched the literature in the PubMed database to verify them.

The first drug we studied was kalamycin. Kalamycin was discovered in 1957 and was used clinically in 1958. Because of its remarkable therapeutic effect on various bacterial infections, especially tuberculosis, it has attracted widespread attention. For the drug kalamycin, we removed 26 lncRNAs related to it and sent the remaining 952 lncRNA candidates to HGNNLDA for prediction. We ranked the candidate 952 lncRNAs in descending order according to the prediction score. The results of the study are shown in Table 1. The results showed that 10 of the top 15 lncRNA candidates were verified by previous literature.

Table 1. The top 15 predicted kalamycin-associated lncRNAs

ncRNA	Pubmed
PDCD4-AS1	Unconfirmed
GS1-24F4.2	Unconfirmed
LOXL1-AS1	32449981
LINC00477	Unconfirmed
RHPN1-AS1	34917132
MCM3AP-AS1	32678686
LINC00937	33376751
EMX2OS	Unconfirmed
MIR210HG	34897892
EXTL3-AS1	Unconfirmed
GATA3-AS1	34358678
ZBED3-AS1	29749482
TOB1-AS1	31482275
MIR31HG	30125988
CACNA1G-AS1	34238292

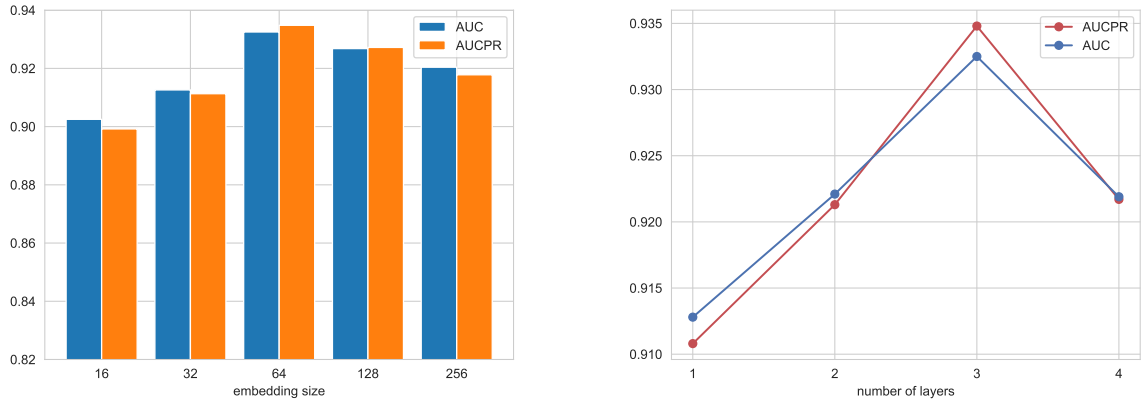
The second drug we studied was bulletyanin. Bulletyanin was discovered in 1985. It has a strong inhibitory effect on mouse sarcoma and mouse liver cancer ascites. For the drug bulletyanin, we first removed 43 lncRNAs associated with it, and the remaining 1772 lncRNA candidates were sent to HGNNLDA for prediction. The top 15 lncRNAs were extracted according to the prediction score. The results are shown in Table 2. The study showed that the previous literature verified 7 of the top 15 lncRNA candidates. It is worth noting that the remaining unverified lncRNA is likely to be associated with drug bulletyanin sensitivity.

Table 2. The top 15 predicted bulletyanin-associated lncRNAs

ncRNA	Pubmed
LINC00710	Unconfirmed
LINC00592	Unconfirmed
LY86-AS1	32854616
THRB-AS1	Unconfirmed
LINC01016	31856144
GS1-24F4.2	Unconfirmed
ERVH48-1	Unconfirmed
WT1-AS	30468780
NR2F1-AS1	34128858
HOXA-AS2	33174119
LOXL1-AS1	32449981
LINC00312	34336850
IGSF11-AS1	Unconfirmed
UBL7-AS1	Unconfirmed
PRKCQ-AS1	Unconfirmed

4 Discussion and Conclusion

Many studies have shown that lncRNA is closely related to drug sensitivity. Given that traditional experimental methods are time-consuming and



(a) The performance of HGNNLDA with different embedding size

(b) The performance of HGNNLDA with different layers

Fig. 3. The performance of HGNNLDA with different parameters

laborious, it is crucial to develop computational methods to predict the association between lncRNA and drug sensitivity. In this work, we develop a computational framework based on a dual-channel hypergraph neural network to predict the association between lncRNA and drug sensitivity, named HGNNLDA. HGNNLDA is the first computational framework to predict the association between lncRNA and drug sensitivity. HGNNLDA constructed lncRNA hypergraph and drug hypergraph, respectively, and used the dual-channel hypergraph neural network to generate lncRNA embedding and drug embedding. HGNNLDA's AUC reached 0.9325, and AUPR reached 0.9348, higher than the other six highly related models. In addition, the two case studies, including drug kalamycin, drug bulletyanin, each with 10, and 7 lncRNA candidates, were verified by previous studies, respectively. The complex experimental results show that HGNNLDA is a reliable tool to infer the association between unknown lncRNA and drug sensitivity. However, there are some limitations that affect the performance of HGNNLDA. The lncRNA-drug sensitivity associations prediction is modeled as a supervised learning task. Due to the limited training data, the quality of learned lncRNA and drug embedding needs further improvement. In the future, we will consider combining self-supervised learning and hypergraph neural networks to improve the performance of lncRNA-drug sensitivity associations prediction.

Acknowledgements

The work was carried out at National Supercomputer Center in Tianjin, and the calculations were performed on Tianhe new generation Supercomputer.

Funding

This work was supported by National Natural Science Foundation of China under grant No. 61972422.

References

Bhat, A. A. *et al.* (2020). Role of non-coding RNA networks in leukemia progression, metastasis and drug resistance. *Mol Cancer*, **19**(1), 57.

Cai, Q. *et al.* (2019). Long non-coding RNA GBCDRlnc1 induces chemoresistance of gallbladder cancer cells by activating autophagy. *Mol Cancer*, **18**(1), 82.

Dayun, L. *et al.* (2021). MGATMDA: Predicting microbe-disease associations via multi-component graph attention network. *IEEE/ACM Trans Comput Biol Bioinform*, **PP**.

Dong, Q. *et al.* (2020). RNAactDrug: a comprehensive database of RNAs associated with drug sensitivity from multi-omics data. *Brief Bioinform*, **21**(6), 2167–2174.

Feng, Y. *et al.* (2018). Hypergraph neural networks. *CoRR*, **abs/1809.09401**.

Geisler, S. and Coller, J. (2013). RNA in unexpected places: long non-coding RNA functions in diverse cellular contexts. *Nat Rev Mol Cell Biol*, **14**(11), 699–712.

Gupta, R. A. *et al.* (2010). Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. *Nature*, **464**(7291), 1071–1076.

Hahne, J. C. and Valeri, N. (2018). Non-Coding RNAs and Resistance to Anticancer Drugs in Gastrointestinal Tumors. *Front Oncol*, **8**, 226.

Hamilton, W. L. *et al.* (2017). Inductive representation learning on large graphs. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 1025–1035.

Harries, L. W. (2012). Long non-coding RNAs and human disease. *Biochem Soc Trans*, **40**(4), 902–906.

Kipf, T. N. and Welling, M. (2016). Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.

Li, G. *et al.* (2018). Predicting microrna-disease associations using label propagation based on linear neighborhood similarity. *Journal of biomedical informatics*, **82**, 169–177.

Li, G. *et al.* (2020). Potential circrna-disease association prediction using deepwalk and network consistency projection. *Journal of Biomedical Informatics*, **112**, 103624.

Liu, D. *et al.* (2021). SMALF: miRNA-disease associations prediction based on stacked autoencoder and XGBoost. *BMC Bioinformatics*, **22**(1), 219.

Liu, M. *et al.* (2020). Predicting miRNA-disease associations using a hybrid feature representation in the heterogeneous network. *BMC Med Genomics*, **13**(Suppl 10), 153.

Lu, C. *et al.* (2021). Deep Matrix Factorization Improves Prediction of Human CircRNA-Disease Associations. *IEEE J Biomed Health Inform*, **25**(3), 891–899.

Lu, M. *et al.* (2008). An analysis of human microrna and disease associations. *PloS one*, **3**(10), e3420.

Luo, J. and Long, Y. (2020). NTSHMDA: Prediction of Human Microbe-Disease Association Based on Random Walk by Integrating Network Topological Similarity. *IEEE/ACM Trans Comput Biol Bioinform*, **17**(4), 1341–1351.

Mercer, T. R. and Mattick, J. S. (2013). Structure and function of long noncoding RNAs in epigenetic regulation. *Nat Struct Mol Biol*, **20**(3), 300–307.

Peng, L. H. *et al.* (2018). Human Microbe-Disease Association Prediction Based on Adaptive Boosting. *Front Microbiol*, **9**, 2440.

Ponting, C. P. *et al.* (2009). Evolution and functions of long noncoding RNAs. *Cell*, **136**(4), 629–641.

Rendle, S. *et al.* (2012). Bpr: Bayesian personalized ranking from implicit feedback. *arXiv preprint arXiv:1205.2618*.

Rinn, J. L. and Chang, H. Y. (2012). Genome regulation by long noncoding RNAs. *Annu Rev Biochem*, **81**, 145–166.

Sharan, R. and Ideker, T. (2008). Protein networks in disease. *Genome Res*, **18**, 644–652.

Shervashidze, N. *et al.* (2011). Weisfeiler-lehman graph kernels. *Journal of Machine Learning Research*, **12**(9).

Sun, J. *et al.* (2014). Inferring novel lncrna–disease associations based on a random walk model of a lncrna functional similarity network. *Molecular BioSystems*, **10**(8), 2074–2081.

Tang, M. *et al.* (2021a). PMDFI: Predicting miRNA-Disease Associations Based on High-Order Feature Interaction. *Front Genet*, **12**, 656107.

Tang, X. *et al.* (2021b). Multi-view Multichannel Attention Graph Convolutional Network for miRNA-disease association prediction. *Brief Bioinform*, **22**(6).

Wang, L. *et al.* (2020). Gcncda: A new method for predicting circrna-disease associations based on graph convolutional network algorithm. *PLOS Computational Biology*, **16**(5), e1007568.

Wei, L. *et al.* (2021). Ganlda: Graph attention network for lncrna-disease associations prediction. *Neurocomputing*.

Wu, Q.-W. *et al.* (2021). Gaerf: predicting lncrna-disease associations by graph auto-encoder and random forest. *Briefings in bioinformatics*.

Xie, X. *et al.* (2020). LncRNA NORAD targets miR-410-3p to regulate drug resistance sensitivity of osteosarcoma. *Cell Mol Biol (Noisy-le-grand)*, **66**(3), 143–148.

Xuan, P. *et al.* (2019). Inferring the disease-associated mirnas based on network representation learning and convolutional neural networks. *International journal of molecular sciences*, **20**(15), 3648.

Xue, H. J. *et al.* (2017). Deep matrix factorization models for recommender systems. In *Twenty-Sixth International Joint Conference on Artificial Intelligence*.

Yang, J. H. *et al.* (2010). deepBase: a database for deeply annotating and mining deep sequencing data. *Nucleic Acids Res*, **38**(Database issue), D123–130.

Yu, Z. *et al.* (2021). Predicting drug-disease associations through layer attention graph convolutional network. *Brief Bioinform*, **22**(4).

Zeng, M. *et al.* (2020). SDLDA: lncRNA-disease association prediction based on singular value decomposition and deep learning. *Methods*, **179**, 73–80.