JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

pubs.acs.org/jcim Article

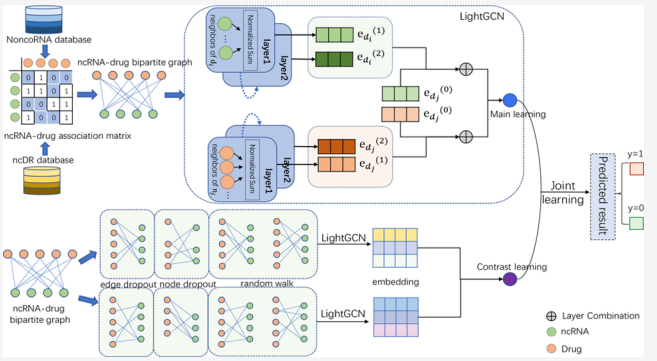
Graph Neural Network with Self-Supervised Learning for Noncoding RNA−Drug Resistance Association Prediction

Jingjing Zheng, Yurong Qian, Jie He, Zerui Kang, and Lei Deng*

Cite This: J. Chem. Inf. Model. 2022, 62, 3676−3684 Read Online

ACCESS | Metrics & More | Article Recommendations

ABSTRACT: Noncoding RNA(ncRNA) is closely related to drug resistance. Identifying the association between ncRNA and drug resistance is of great significance for drug development. Methods based on biological experiments are often time-consuming and small-scale. Therefore, developing computational methods to distinguish the association between ncRNA and drug resistance is urgent. We develop a computational framework called GSLRDA to predict the association between ncRNA and drug resistance in this work. First, the known ncRNA−drug resistance associations are modeled as a bipartite graph of ncRNA and drug. Then, GSLRDA uses the light graph convolutional network (lightGCN) to learn the vector representation of ncRNA and drug from the ncRNA−drug bipartite graph. In addition, GSLRDA uses different data augmentation methods to generate different views for ncRNA and drug nodes and performs self-supervised learning, further improving the quality of learned ncRNA and drug vector representations through contrastive learning between nodes. Finally, GSLRDA uses the inner product to predict the association between ncRNA and drug resistance. To the best of our knowledge, GSLRDA is the first to apply self-supervised learning in association prediction tasks in the field of bioinformatics. The experimental results show that GSLRDA takes an AUC value of 0.9101, higher than the other eight state-of-the-art models. In addition, case studies including two drugs further illustrate the effectiveness of GSLRDA in predicting the association between ncRNA and drug resistance. The code and data sets of GSLRDA are available at https://github.com/JJZ-code/GSLRDA.



INTRODUCTION

Noncoding RNAs (ncRNAs) are DNA transcripts that cannot be encoded into proteins.[1] Numerous studies have shown that ncRNAs are involved in many biological functions, such as cell proliferation, cell cycle progression, and apoptosis.[2−4] They have been shown to be key regulators of gene expression, not just "byproducts" of gene transcription.[5,6] In recent years, noncoding RNAs such as long noncoding RNAs (lncRNAs), microRNAs (miRNAs), and circular RNAs (circRNAs) have received extensive attention. lncRNAs are antisense RNA molecules that can specifically bind to noncoding regions of target genes, regulate gene transcription and expression, and play a role in promoting or suppressing tumors.[7] miRNAs are a class of noncoding RNAs that regulate gene transcription and expression and participate in a variety of physiological activities.[8] circRNAs have stability, making them stable in plasma, saliva, and other peripheral tissues, and are novel biomarkers and new targets for cancer diagnosis and treatment.[9,10]

Cancer is one of the major diseases that seriously endangers human health. Currently, the leading cancer treatments are surgery, radiotherapy, and chemotherapy.[11] Chemotherapy is one of the essential treatment methods for cancer at present. Still, most patients often develop drug resistance during chemotherapy, which leads to the recurrence and metastasis of cancer cells[12] and the failure of cancer treatment. Exploring the molecular mechanisms of drug resistance is crucial for drug discovery and cancer treatment.

With the development of sequencing technology, experimental studies have found that ncRNAs are closely related to many diseases, including malignant tumors. Studies have shown that the lncRNA NKILA enhances the sensitivity of T cells to activation-induced cell death by inhibiting NF-κB signaling in breast cancer and the lung cancer microenvironment, thus promoting the immune escape of non-small-cell lung cancer (NSCLC) cells and affecting the immune tolerance of lung cancer.[13] miRNAs such as miR-140-5p and miR-146a can also play an essential role in doxorubicin-

Received: March 29, 2022
Published: July 15, 2022

https://doi.org/10.1021/acs.jcim.2c00367
J. Chem. Inf. Model. 2022, 62, 3676−3684

induced cardiotoxicity by targeting Sirt2, Nrf2, TAF9b/P53, and other pathways.[14,15] CircPAN3 is a critical mediator in the development of resistance in acute myeloid leukemia (AML). Experiments show that CircPAN3 promotes drug resistance in AML by regulating protein expression.[16] Abnormal expression of ncRNAs can regulate tumor drug resistance, which provides new opportunities and research directions for overcoming tumor drug resistance. Traditional biological experiments often consume a lot of material and financial resources, which makes them difficult to implement to a certain extent. Computational methods are undoubtedly useful accelerators for this process, and few computational methods have explored the relationship between ncRNAs and drug resistance. Li et al.[17] developed the method (LRGCPND) of the graph neural network to efficiently identify potential ncRNA−drug resistance associations. This is the only existing ncRNA−drug resistance association prediction using a computational approach. In their studies, first, neighbor information on nodes in the ncRNA−resistance bipartite graph is captured by aggregation, then feature transformation is performed by linear operations. Finally, they use residual blocks to fuse the features of low-level nodes to achieve prediction.

Although there are few computational methods for predicting ncRNA−drug resistance, many related association prediction methods are worth discussing. Dayun et al.[18] proposed a novel computational framework of MGATMDA to detect microbial−disease associations by multicomponent graph attention networks. First, they generated the latent vectors of nodes from the bipartite graph through the decomposer. Then they obtained the unified embedding representation through the combinator. Finally, they used the attention mechanism for microbial−disease associations prediction. Ji et al.[19] constructed a GATNNCDA model combining graph neural network and multilayer neural network for predicting potential associations of circRNA−disease. Deng et al.[20] proposed a new method (Graph2MDA) using variational graph autoencoders to predict microbe−drug associations. Graph2MDA first constructs a multimodal attribute graph of microbe and drugs, then uses a variational graph autoencoder (VGAE) to learn the latent representations of nodes. Finally, a deep neural network is used to predict potential microbe−drug associations. Inspired by the heterogeneous attention network (HGAT), Zhao et al.[21] developed a new heterogeneous attention network framework, HGATLDA, based on metapaths, which is used to predict the relationship between lncRNAs and diseases. Fan et al.[22] proposed a new prediction method based on graph convolution matrix completion, GCRFLDA. GCRFLDA embeds the conditional random field (CRF) with an attention mechanism into the coding layer, preserves the similarity information between graph nodes, and scores the potential lncRNA−disease association. Wang et al.[23] constructed a computing method, GCNCDA, based on deep learning, fast learning, and the graph convolution network (FastGCN) algorithm. GCNCDA used a forest penalty attribute classifier to predict potential associations and diseases between circRNAs accurately. Lan et al.[24] proposed a new computing framework (IGNSCDA) based on the improved graph convolution network and negative sampling to infer the association between circRNAs and disease. Li et al.[25] proposed a SDNE-MDA model based on structured deep network embedding (SDNE) to predict miRNA−disease associations (MDAs). The model constructs a complex network molecular association network (MAN) by combining miRNA, disease, and three related molecules (lncRNA, drug, protein) and their relationships.

Although the association prediction approaches such as graph neural network technology have been widely used in various fields, there are still some limitations in the field of ncRNA−drug resistance association prediction. Most models tackle the task of association prediction based on supervised learning. These supervised signals come from the observed ncRNA−drug resistance associations; however, the observed associations are very sparse. In this work, we developed a computational model called GSLRDA to infer unknown ncRNA−drug resistance associations. GSLRDA combines graph neural networks and self-supervised learning. GSLRDA designs the main task and auxiliary tasks. In the main task, GSLRDA takes an ncRNA−drug resistance bipartite graph as input, uses a lightGCN to learn ncRNA and drug vector representations, and then uses the inner product to predict ncRNA and drug resistance. In the auxiliary task, GSLRDA first generates different perspectives for ncRNA and drug nodes through different data augmentation methods and then performs comparative learning between nodes to improve the quality of the learned ncRNA and drug vector representation. Complicated experimental results show that GSLRDA is superior to the existing eight excellent calculation methods. Ablation experiments verify the effectiveness of self-supervised learning. In addition, the case study results on two drugs indicate that GSLRDA is an effective tool for predicting the association between ncRNA and drug resistance.

METHODS

Data Set. Known ncRNA−drug resistance association pairs come from the NoncoRNA and ncDR databases. NoncoRNA[26] (http://www.ncdtcdb.cn:8080/NoncoRNA) is the first database that provides experimentally supported associations between 5568 ncRNAs, 154 drugs, and 134 cancers.

ncDR[27] (http://www.jianglab.cn/ncDR) is a noncoding RNA (ncRNA) database. The ncDR database contains 5864 associations between 877 miRNAs, 162 lncRNAs (1039 ncRNAs in total), and 145 compounds obtained from 900 published articles.

We removed redundant data from the NoncoRNA and ncDR databases; 2693 known ncRNA−drug resistance-associated pairs were obtained, including 121 drugs and 625 ncRNAs. In our base data set, there are only 2693 ncRNA−drug resistance associations. To avoid the effect of sample imbalance, we randomly sample 2693 negative samples from unknown associations to achieve the same number as positive samples. In addition, the independent data set was created by searching the PubMed database literature. It contains 534 known ncRNA resistance associations, including 168 ncRNAs and 70 drugs.

ncRNA−Drug Bipartite Graph. The bipartite graph[28] abstracts the relationship between ncRNA and drug resistance as a graph. Let N and D be the set of ncRNAs and drugs, respectively. Let $E = \{(y_{nd} | n \in N, d \in D)\}$ indicate the verified association between ncRNA and drug resistance. We use the ncRNA−drug resistance association matrix A to construct a bipartite graph $G = (V, E)$, where the node set V contains all ncRNAs and drugs, $V = N \cup D$.

GSLRDA. GSLRDA takes the bipartite graph of ncRNA and drugs as input and outputs potential ncRNA and drug resistance associations. GSLRDA uses lightGCN to learn the representation of ncRNAs and drugs from the bipartite graph

https://doi.org/10.1021/acs.jcim.2c00367
J. Chem. Inf. Model. 2022, 62, 3676−3684

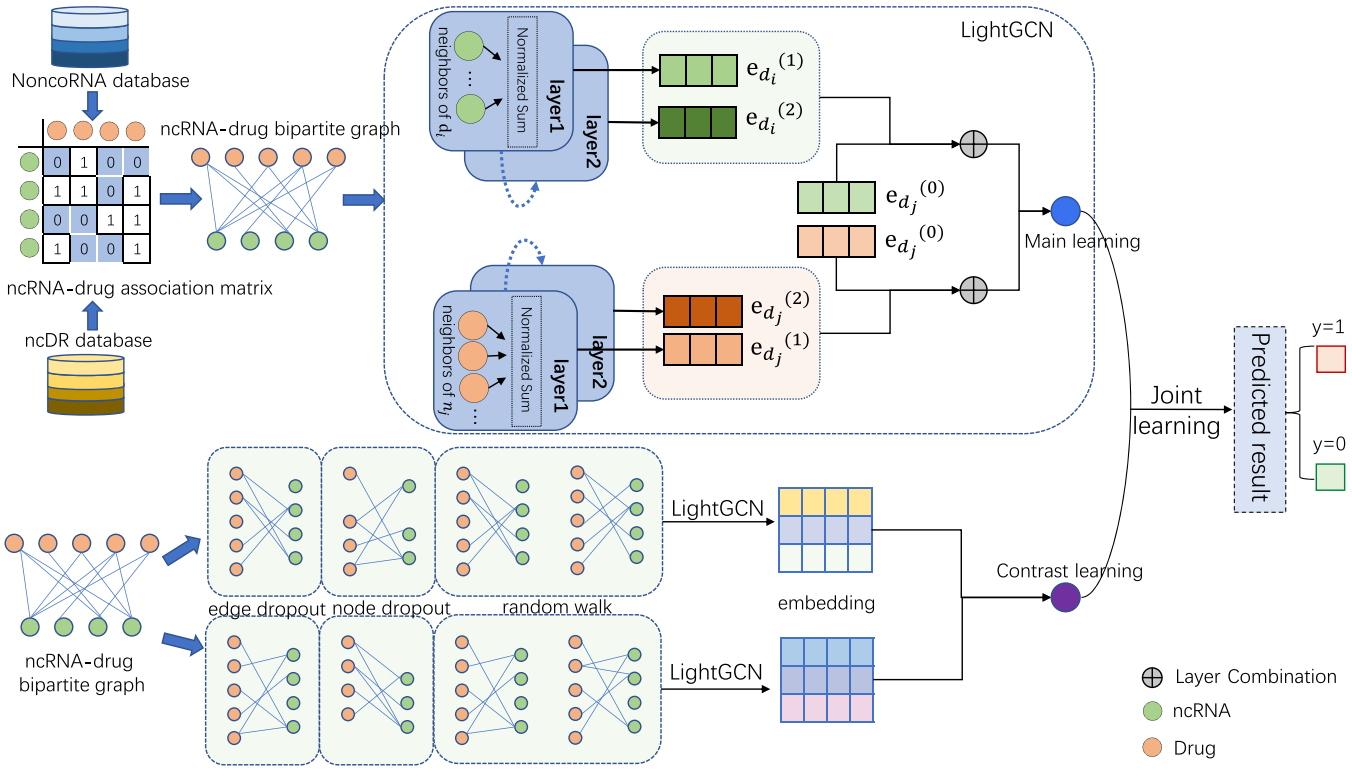


Figure 1. Overview of GSLRDA. GSLRDA designs two tasks. In the main task, we first model the known association between ncRNA and drug resistance into a bipartite graph of ncRNA and drug. Then, the bipartite graph of ncRNA and drug is input into lightGCN to learn the representation of ncRNA and drug node. Finally, we use inner products to infer the association between unknown ncRNAs and drug resistance. Due to the sparse supervision signal, the quality of the learned ncRNA and drug representation needs to be further improved. Therefore, we perform self-supervised learning by designing auxiliary tasks. In the auxiliary task, we generate different views for ncRNA and drug nodes through three data augmentation methods: node loss, edge loss, and random walk. Then, we perform comparative learning between nodes. Finally, we integrate the main task and auxiliary task, and jointly optimize the loss function, to predict the associations of ncRNA−drug resistance.

of ncRNAs and drugs and uses the inner product to infer the relationship between ncRNAs and drugs. Due to the sparse supervision signal, the learned ncRNA and drug representations are insufficient. To further improve the learned association of ncRNAs and drug resistance and improve the model's prediction performance, we designed auxiliary tasks by performing self-supervised learning. Specifically, first, we use different data augmentation methods to generate other views for ncRNA and drug nodes and then perform contrastive learning tasks between nodes. Finally, we optimize the ncRNA−drug association prediction and comparative learning tasks jointly. The GSLRDA model is shown in Figure 1.

GCN for ncRNA−Drug Resistance Association. Graph embedding represents from a single ID to high-order neighbors, which makes the graph neural network (GCN) successful in the recommendation task. To better learn the high-level features of ncRNAs and drugs, we use the advanced lightweight graph neural network, lightGCN.

Initialization: Each ncRNA and drug is associated with an ID embedding. Here we use e_n^0 and e_d^0 to represent the ID embedding of ncRNA n and drug d, respectively.

Simplifying and powering graph convolution: A basic idea of a graph neural network is to embed the target node based on its neighbor information. Intuitively, the information on each node and its surrounding nodes is aggregated through a neural network. The neighborhood aggregation formula of ncRNAs and drugs is

$$e_n^{(k+1)} = F_{\text{agg}}(e_n^{(k)}, \{e_d^{(k)}: d \in D_n\}) \tag{1}$$

$$e_d^{(k+1)} = F_{\text{agg}}(e_d^{(k)}, \{e_n^{(k)}: n \in N_d\}) \tag{2}$$

where D_n represents the set of drugs that have interacted with ncRNA n, and N_d represents the embedding representation of ncRNA and drug after k-layer propagation. When $k = 0$, e_n^0 and e_d^0 represent ncRNA n and drug d, respectively. F_{agg} is an aggregate function.

In graph convolution operations, the aggregation function F_{agg} is the core of the operation. LightGCN abandons the feature transformation and nonlinear activation operations that have no positive significance to the model performance and only uses simple weighted summation operations. In the lightGCN model, the calculation formula for the neighborhood aggregation of ncRNA and drug nodes is

$$e_n^{(k+1)} = \sum_{d \in D_n} \frac{1}{\sqrt{|D_n|}\sqrt{|N_d|}} e_d^{(k)} \tag{3}$$

$$e_d^{(k+1)} = \sum_{n \in N_d} \frac{1}{\sqrt{|N_d|}\sqrt{|D_n|}} e_n^{(k)} \tag{4}$$

where $\frac{1}{\sqrt{|N_d|}\sqrt{|D_n|}}$ is the normalization operation.

Layer combination: After the k-layer graph convolution operation, we combine the embeddings obtained from each layer and then form the final representation of ncRNAs and drugs. The formulas are

https://doi.org/10.1021/acs.jcim.2c00367
J. Chem. Inf. Model. 2022, 62, 3676−3684

$$e_n = \sum_{k=0}^{K} \beta_k e_n^{(k)} \tag{5}$$

$$e_d = \sum_{k=0}^{K} \beta_k e_d^{(k)} \tag{6}$$

In these equations, β_k is the hyperparameter and is set to $\frac{1}{K+1}$.

Prediction layer: The inner product of an ncRNA and the final representation of the drug are used as a model prediction.

$$y'_{nd} = e_n^{\mathrm{T}} e_d \tag{7}$$

Loss function: We adopt Bayesian Personalized Ranking (BPR) loss as the loss function.

$$L' = -\sum_{n=1}^{M} \sum_{d \in D_n} \sum_{q \notin D_n} \log \sigma(y'_{nd} - y'_{nq}) \tag{8}$$

Self-Supervised Learning. To further improve the prediction performance of the model, we design auxiliary tasks by performing self-supervised learning. Self-supervised learning in ncRNA and drug resistance association prediction includes two parts: data enhancement and comparative learning. The detailed process is as follows.

Data enhancement: There is an inherent link between ncRNAs and drug resistance and they do not exist independently. Methods such as NLP (synonym substitution, random deletion, etc.) and CV (flip, Gaussian white noise, etc.) tasks to achieve data enhancement are not suitable for ncRNA−drug resistance association prediction. Therefore, it is necessary to develop new enhanced algorithms for the prediction of ncRNA and drug resistance associations.

We use three algorithms, node dropout, edge dropout, and random walk, in the graph structure to obtain different views of ncRNA and drug nodes:

Node dropout deletes nodes and their linked edges in the graph through probability ρ, $\rho \in (0, 1)$.

$$\hat{G} = (V \odot O, E); \quad O = (O_1, O_2, \ldots O_i \ldots, O_n); \quad O_i \approx \rho \tag{9}$$

O is a vector responsible for deciding which nodes in the node set V should be retained.

Edge dropout deletes the edges in the original graph with probability ρ, $\rho \in (0, 1)$.

$$\hat{G} = (V, E \odot Q); \quad Q = (Q_1, Q_2, \ldots Q_i \ldots, Q_n); \quad Q_i \approx \rho \tag{10}$$

Q is a vector that is responsible for deciding which edges in the edge set E should be deleted.

Both node dropout and edge dropout generate shared subgraphs between graph convolutional layers. We consider using random walk operators to allocate different subgraphs to different layers.

$$\hat{G} = (V, E \odot Q'); \quad Q' = (Q'_1, Q'_2, \ldots, Q'_n); \quad Q'_i \approx \rho \tag{11}$$

Q' is a vector that is responsible for deciding which edges in the edge set E should be deleted.

After obtaining graph structures from different perspectives through the above three data enhancement methods, the lightGCN in the previous section is used to learn node features.

Comparative learning: For ncRNA n, $n \in N$, we define n^+ as a positive sample similar to n, n^- as a negative sample not similar to n, and s as a metric function to measure the similarity between samples. The goal of contrastive learning is to learn an encoder f such that $S(f(n), f(n^+)) \gg S(f(n), f(n^-))$. We use the vector product to calculate the similarity between two samples; then, the contrastive learning loss function of the ncRNA node is

$$L''_n = -E_N\left[\log \frac{\exp(f(n)^{\mathrm{T}} f(n^-))}{\exp(f(n)^{\mathrm{T}} f(n^+)) + \sum_{j=1}^{N-1} \exp(f(n)^{\mathrm{T}} f(n_j^-))}\right] \tag{12}$$

n has a positive sample and $N-1$ negative samples, and the goal of our learning is to make the features of n more similar to the features of n^+ and less similar to the features of the $N-1$ negative samples. Similarly, we can obtain the loss function L''_d of the drug node; then, the loss function of contrastive learning is $L'' = L''_n + L''_d$.

Joint Learning. To further improve the performance of the model, we jointly optimize the lightGCN and self-supervised learning tasks using a multitask training strategy.

$$L = L' + \gamma_1 L'' + \gamma_2 \|\theta\|_2^2 \tag{13}$$

where θ is the parameter set of the lightGCN model, since no additional parameters are introduced in self-supervised learning; γ_1 and γ_2 are hyperparameters that control the strength of self-supervised learning and L2 regularization, respectively.

EXPERIMENTS AND RESULTS

Experimental Setup. To evaluate the performance of the GSLRDA model, a 5-fold cross-validation method was used to evaluate the potential of ncRNAs to predict drug resistance. The 2693 known ncRNA−resistance-related data are randomly divided into 5 subsets of the same size. The classification criteria are as follows:

$$P = P_1 \cup P_2 \cup P_3 \cup P_4 \cup P_5 \tag{14}$$

$$\varnothing = P_1 \cap P_2 \cap P_3 \cap P_4 \cap P_5 \tag{15}$$

Each time the model is trained and evaluated, the current target subset is used as the test set, and the remaining 4 subsets are used as the training set. This process continues until each subset is used as the test set. Then, we calculate the average value of the 5 iterations as the final result of the GSLRDA model. We choose the commonly used evaluation indicators in association prediction tasks, including AUC (area under the receiver operating characteristic curve)[29] and AUPR (area under the accurate recall curve).[30]

Comparison of Models. To prove the effectiveness of the GSLRDA model, we compared it with 8 methods of making association predictions on biological information.

LRGCPND. The first computational model to predict ncRNA resistance, LRGCPND[17] captures the neighbor information representation in the bipartite graph of ncRNA resistance through aggregation, then performs feature transformation through linear operations, and finally makes the final prediction through residual links.

SDLDA. SDLDA[31] is calculation method for predicting lncRNA-disease by combining the nonlinear features and linear

https://doi.org/10.1021/acs.jcim.2c00367
J. Chem. Inf. Model. 2022, 62, 3676−3684

features obtained in deep learning and singular value matrix decomposition.

DMFCDA. In DMFCDA,[32] a method of deep matrix decomposition, using a projection layer composed of a fully connected network to capture the potential characteristics of circRNA and diseases, the combination is sent to a multilayer neural network for prediction.

DMFMDA. In DMFMDA,[33] the one-hot encoding of microbe and disease is input to the embedding layer to convert it into a low-dimensional vector. Then, the matrix decomposition is realized through the neural network with the embedded layer, and finally, the prediction is made.

KATZHMDA. In KATZHMDA,[34] the heterogeneous network is constructed from the multisource similarity network of miRNA and disease and the miRNA−disease association network, and finally, the miRNA−disease association is predicted by KATZ.

NTSHMDA. In NTSHMDA,[35] a heterogeneous microbial−disease network is constructed and then microbial−disease associations are predicted through an integrated network based on random walks.

AE-RF. For AE-RF,[36] the deep features of circRNA and disease are extracted through a deep autoencoder, and random forest is used to make association predictions.

ABHMDA. With ABHMDA,[37] first, the similarity between diseases and microorganisms is calculated, and then reliable negative samples are selected through K-means clustering. Finally, the strong classification adaptive boosting combined by multiple weak classifiers predicts the human microbe−disease association.

In this work, we used 5-fold cross-validation to evaluate the performance of GSLRDA and the other 8 methods. As shown in Figure 2, the AUC value reached 0.9101. Overall, our

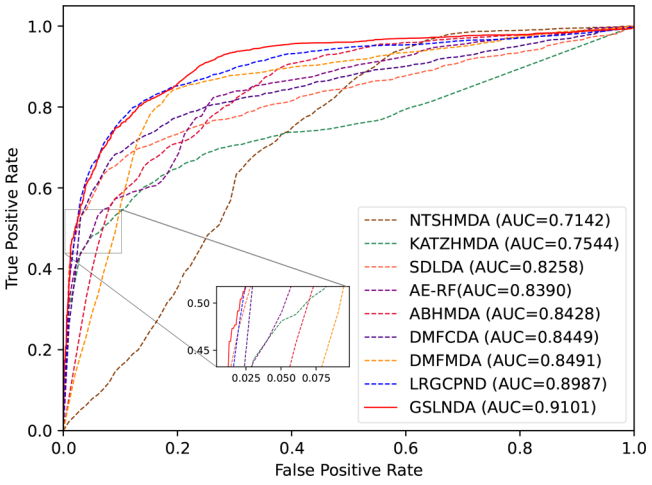


Figure 2. ROC of GSLRDA compared with eight related models.

method outperformed other association prediction methods. This may be attributed to two strategies, self-supervised learning and the graph neural network, which enable GSLRDA to capture richer and more important feature information. Compared with matrix factorization methods (SDLDA,[31] DMFCDA,[32] DMFMDA[33]), we deepened the learning from the representation of matrix factorization to the use of graph neural networks to capture richer information using the high-order connectivity of ncRNAs and drug resistance. Compared with supervised learning methods (LRGCPND,[17] KATZHMDA,[34] NTSHMDA,[35] AE-RF,[36] ABHMDA[37]), we used self-supervised learning node enhancement to construct a comparative learning strategy, obtained more important information from the data, and further improved the model performance. Overall, GSLRDA is effective in predicting the association of ncRNA resistance.

To further verify the predictive ability of the GSLRDA model, we established an independent test set and compared GSLRDA with other excellent models on the independent test set. Through a literature search in the PubMed database, an independent test set containing 534 ncRNA and drug resistance associations, 168 ncRNAs, and 70 drugs was established. We used the 526 ncRNA and drug resistance associations of our data set as the training set training model and tested it on the independent test set. The experimental results are shown in Figure 3. The AUC of GSLRDA reaches

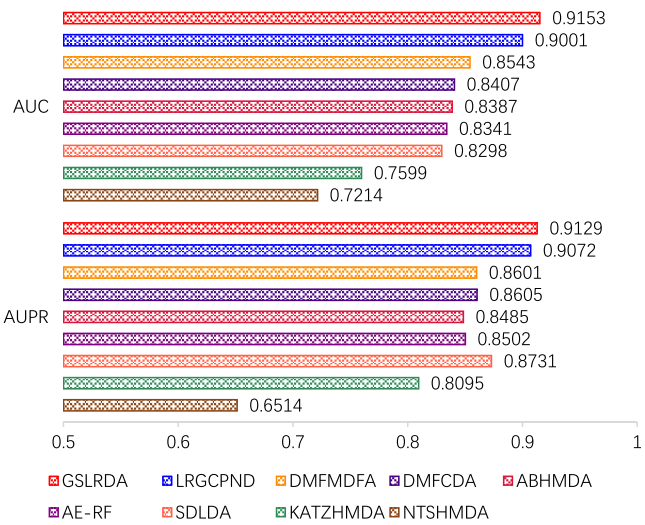


Figure 3. Performance compared between GSLRDA and eight related models on independent test sets.

0.9153, and the AUPR reached 0.9129, both higher than those of the other models. The experimental results fully show that GSLRDA is an effective tool to infer the association between ncRNA and drug resistance.

Ablation Experiment. In this work, we designed a GSLRDA model that uses a self-supervised learning mechanism to assist lightGCN[38] in predicting the association between ncRNAs and drug resistance. To analyze the necessity of the self-supervised learning strategy for the GSLRDA model, we conducted ablation experiments on this. The ultimate goal of our task is to predict whether there is an association between ncRNAs and drugs, that is, to provide useful ncRNA targets for drugs. This is analogous to the popular recommendation task. Therefore, we choose the NGCF[39] model with the classical GCN method to prove the necessity of introducing self-supervised learning. Specifically, different models were used to predict ncRNA−resistance associations on the same data set. Table 1 shows the average evaluation metric values obtained for the three models under 5-fold cross-validation. GSLRDA consistently outperforms other baseline methods. This verifies the rationality and effectiveness of introducing self-supervised learning. The lightGCN implementation performs better than NGCF, which is a consistent claim of the lightGCN paper. For the GSLRDA model, using self-

https://doi.org/10.1021/acs.jcim.2c00367
J. Chem. Inf. Model. 2022, 62, 3676−3684

Table 1. Performance Comparison between GSLRDA, LightGCN, and NGCF

methods	GSLRDA	LightGCN	NGCF
AUC	0.9101	0.8858	0.8769
AUPR	0.9144	0.8905	0.8530

supervised learning to assist in predicting the performance of ncRNA−resistance associations, the AUC value increased by 2.4−3.3% compared to the model using only the GCN method. This further indicates that GSLRDA has an important guiding role in the discovery of drug resistance-related ncRNAs.

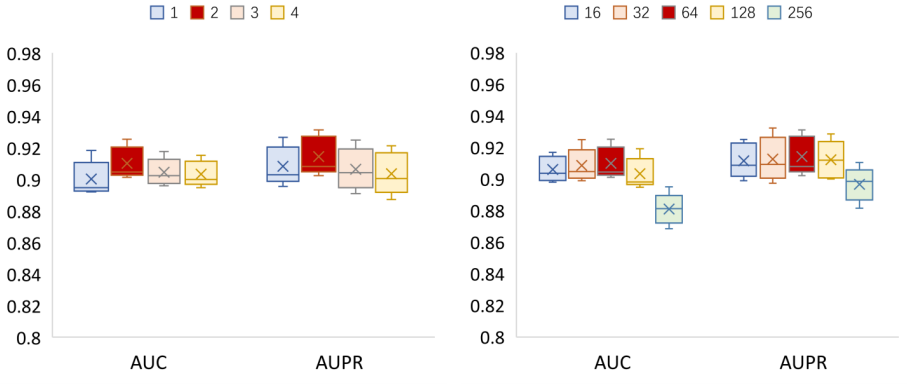
Influence of Parameters. In this work, we evaluated the influence of parameters on the performance of the GSLRDA model. The influence of two important parameters, the number of GCN layers and embedding sizes were introduced. We changed one of the parameters, kept the other parameters unchanged, and performed 5-fold cross-validation.

Effect of GCN Layers. The rest of the parameters remained unchanged, the GCN layers were selected from 1, 2, 3, or 4 to change in turn, and the 5-fold cross-validation was used for evaluation. The AUC and AUPR values under 5-fold cross-validation can be found in Figure 4a. When GCN layer = 2, the performance of the GSLRDA model was optimal. As the number of layers increased, the performance gradually decreased. The increase in the number of GCN layers caused the learned feature vectors to be smooth and lose important information. It can be seen from the results that setting the GCN layer to 2 can solve the smoothing problem very well.

Effect of Embedding Sizes. The other parameters remained unchanged, embedding sizes were chosen from 16, 32, 64, 128, or 256, and 5-fold cross-validation was performed. The AUC and AUPR values under 5-fold cross-validation can be found in Figure 4b. When the embedding size increased, the performance of the GSLRDA model also increased until embedding size = 64, and the model reached the optimum.

Embedding Visualization. To further explain the learning ability of the GSLRDA model, we visualized the ncRNA and drug features. Specifically, we first constructed 625 ncRNAs and 121 drugs into 652 × 121 = 75 625 ncRNA−drug pairs. There were 2693 known ncRNA−drug resistance association pairs, and the rest were unknown association pairs. Then, we used t-SNE[40] to visualize the features of ncRNA−drug pairs. t-SNE is a technology that integrates dimensionality reduction and visualization, which can project high-dimensional feature vectors into a 2-dimensional or 3-dimensional space. Figure 5a,b projects 75 625 ncRNA−drug pairs embedded into 2D space. Blue + represents an unknown ncRNA−drug resistance pair, and an orange dot represents the known associated pair. Figure 5a shows the embedding of the initial ncRNA−drug resistance associations. Figure 5b is the embedding of the ncRNA−drug resistance associations after learning by the GSLRDA model. Comparing Figure 5a,b, we can see that the GSLRDA model can better aggregate known association pairs, making it easier to distinguish them from unknown association pairs. In addition, we visualized the learned drug embedding and ncRNA embedding, as shown in Figure 5c,d, respectively. Figure 5c shows the embedding visualization of the drug node after model learning. The drugs imatinib and etoposide are associated with four ncRNAs of the same type. The drugs imatinib and trastuzumab are associated with only one ncRNA of the same type. Therefore, the drugs imatinib and etoposide are more similar. From Figure 5c, we can see that the distance between imatinib and etoposide is smaller. Figure 5d shows the visualization of ncRNA node embedding after model learning. ncRNA circPVT1 and GAS5 have two identical drug associations, while circPVT1 and circBA9.3 do not have the same drug association. In Figure 5d, we can see that ncRNA circPVT1 and GAS5 are more similar. Experimental results show that our model can effectively learn the potential features of ncRNAs and drugs.

Case Study. In this section, we conduct a case study to demonstrate the effectiveness of GSLRDA in predicting a new association between ncRNA and drug resistance. Temozolomide[41] and 5-fluorouracil[42] (5-FU) were selected and studied. The most widely used drug in glioblastoma[43] (GBM) treatment is temozolomide. More than half of patients develop resistance to temozolomide and fail treatment. 5-FU is often used to treat colorectal cancer[44] (CRC). The human body's resistance to 5-FU is a major obstacle to the treatment of CRC. Therefore, the discovery of ncRNA related to drug resistance plays a positive role in disease treatment. For each drug, first, we remove the ncRNA related to the drug in the data set and treat it as a new drug. GSLRDA was implemented to predict and sort the ncRNA of drugs in descending order according to



(a) Effect of GCN layer (b) Effect of embedding size

Figure 4. Box plot of the influence of the parameter (the cross indicates the mean value).

https://doi.org/10.1021/acs.jcim.2c00367
J. Chem. Inf. Model. 2022, 62, 3676−3684

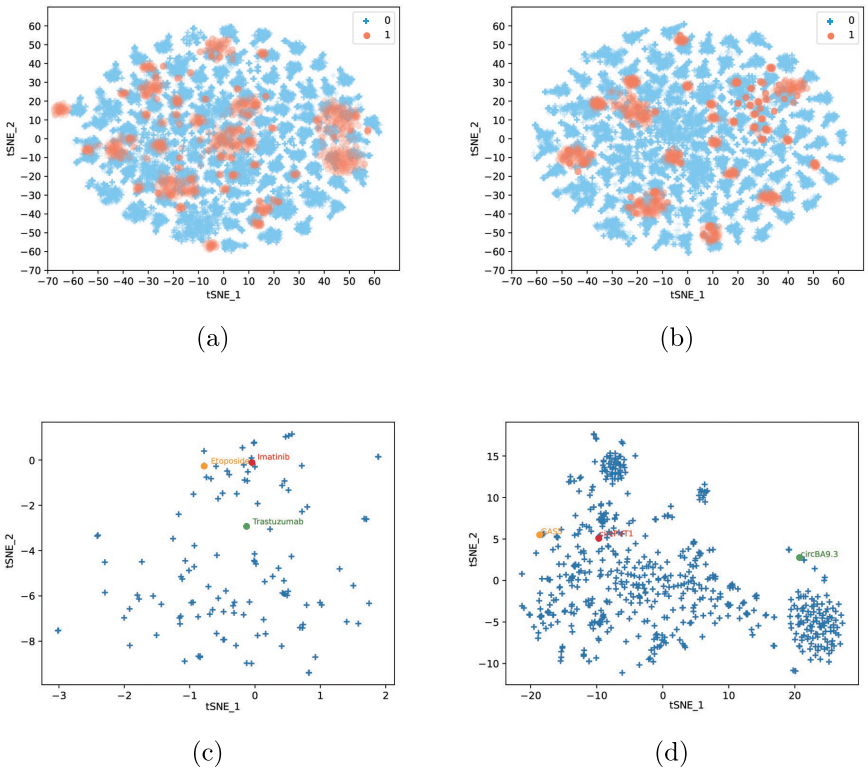


Figure 5. Embedding visualization: (a) embedding of the initial ncRNA and drug resistance associated pair; (b) embedding of the ncRNA and drug resistance associated pair learned by the GSLRDA model; (c) embedding of the drug node; (d) embedding of the ncRNA node.

the association score. We confirmed the top 10 ncRNAs in PubMed. The results in Tables 2 and 3 show that more than half of the ncRNAs of the two drugs are sufficiently proven, which also shows that the GSLRDA model is good for predicting the relationship between ncRNA and drug resistance. It is worth noting that there is a possibility of a high correlation between unproven ncRNA and drugs, which is worthy of further study.

Table 2. Top 10 miRNAs Related to Temozolomide Resistance Predicted by GSLRDA

ncRNA	PubMed ID
miR-181a	31190889
miR-146a	32973101
miR-200c	34245265
miR-30b	33408780
miR-26b	28898169
miR-99a	unconfirmed
miR-155	32220051
miR-34a	33765907
miR-193b	unconfirmed
miR-210	31190889

Table 3. Top 10 miRNAs Related to 5-Fluorouracil Resistance Predicted by GSLRDA

ncRNA	PubMed ID
miR-99a-5p	unconfirmed
miR-21	33569416
miR-125b-5p	32649737
XIST	30907503
Let-7	29330293
miR-Plus-A1031	unconfirmed
miR-146b-5p	unconfirmed
miR-191*	29737579
CASC11	unconfirmed
miR-1256	unconfirmed

DISCUSSION AND CONCLUSION

A large number of studies have shown that ncRNAs play a vital role in drug resistance. Identifying the association between ncRNAs and drug resistance is of great significance for developing drugs and conducting clinical trials. Predicting the association between ncRNAs and drug resistance based on computational methods is convenient and large-scale. The existing methods suffer from sparse supervision signals. In this work, we proposed a method called GSLRDA, which combines graph neural networks and self-supervised learning. GSLRDA uses lightGCN to learn vector representations of ncRNAs and drugs and uses self-supervised learning by designing auxiliary tasks to improve the quality of the learned vector representations of ncRNAs and drugs. The experimental results show that GSLRDA had the best AUC value, 0.9101, compared with the other excellent models. The results of ablation experiments show that the application of self-supervised learning can indeed further improve the prediction effect of the model. In addition, case studies on two drugs were performed. Among the predicted top ten candidate ncRNAs, temozolomide and 5-fluorouracil drugs have 9 and 7 associated ncRNAs, respectively, which have been validated by previous studies. The complex experimental results show that GSLRDA is a reliable predictor of potential ncRNA and drug resistance. Although the use of self-supervised learning alleviates the

https://doi.org/10.1021/acs.jcim.2c00367
J. Chem. Inf. Model. 2022, 62, 3676−3684

problem of data sparseness, there are still few known ncRNA−drug resistance associations. In future work, we will collect more ncRNA−drug resistance associations and adopt more data augmentation methods to explore the information of graphs better and enhance the performance. In addition, we will also pay more attention to the pretraining of tasks to improve the transferability of the model.

DATA AND SOFTWARE AVAILABILITY

The code and data sets of GSLRDA are available at https://github.com/JJZ-code/GSLRDA.

AUTHOR INFORMATION

Corresponding Author

Lei Deng − *School of Software, Xinjiang University, Urumqi 830091, China; School of Computer Science and Engineering, Central South University, Changsha 410083, China*; Email: leideng@csu.edu.cn

Authors

Jingjing Zheng − *School of Software, Xinjiang University, Urumqi 830091, China*; orcid.org/0000-0002-8549-4235

Yurong Qian − *School of Software, Xinjiang University, Urumqi 830091, China*

Jie He − *School of Computer Science and Engineering, Central South University, Changsha 410083, China*

Zerui Kang − *School of Computer Science and Engineering, Central South University, Changsha 410083, China*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jcim.2c00367

Funding

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

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

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

REFERENCES

(1) Slack, F. J.; Chinnaiyan, A. M. The Role of Non-coding RNAs in Oncology. *Cell* **2019**, *179*, 1033−1055.

(2) Ferreira, D.; Escudeiro, A.; Adega, F.; Anjo, S. I.; Manadas, B.; Chaves, R. FA-SAT ncRNA interacts with PKM2 protein: Depletion of this complex induces a switch from cell proliferation to apoptosis. *Cell. Mol. Life Sci.* **2020**, *77*, 1371−1386.

(3) Anastasiadou, E.; Jacob, L. S.; Slack, F. J. Non-coding RNA networks in cancer. *Nature Reviews Cancer* **2018**, *18*, 5−18.

(4) Liu, Y.; Liu, X.; Lin, C.; Jia, X.; Zhu, H.; Song, J.; Zhang, Y. Noncoding RNAs regulate alternative splicing in Cancer. *J Exp Clin Cancer Res.* **2021**, *40*, 11.

(5) Cech, T. R.; Steitz, J. A. The noncoding RNA revolution-trashing old rules to forge new ones. *Cell* **2014**, *157*, 77−94.

(6) Dhamija, S.; Diederichs, S. From junk to master regulators of invasion: lncRNA functions in migration, EMT and metastasis. *International journal of cancer* **2016**, *139*, 269−280.

(7) Hardin, H.; Helein, H.; Meyer, K.; Robertson, S.; Zhang, R.; Zhong, W.; Lloyd, R. V. Thyroid cancer stem-like cell exosomes: regulation of EMT via transfer of lncRNAs. *Laboratory Investigation* **2018**, *98*, 1133−1142.

(8) Nunez Lopez, Y. O. N.; Victoria, B.; Golusinski, P.; Golusinski, W.; Masternak, M. M. Characteristic miRNA expression signature and random forest survival analysis identify potential cancer-driving miRNAs in a broad range of head and neck squamous cell carcinoma subtypes. *Reports of Practical Oncology and Radiotherapy* **2018**, *23*, 6−20.

(9) Chen, B.; Huang, S. Circular RNA: an emerging non-coding RNA as a regulator and biomarker in cancer. *Cancer letters* **2018**, *418*, 41−50.

(10) Ojha, R.; Nandani, R.; Chatterjee, N.; Prajapati, V. K. Emerging role of circular RNAs as potential biomarkers for the diagnosis of human diseases. *Adv Exp Med Biol.* **2018**, *1087*, 141−157.

(11) Siegel, R. L.; Miller, K. D.; Sauer, A. G.; Fedewa, S. A.; Butterly, L. F.; Anderson, J. C.; Cercek, A.; Smith, R. A.; Jemal, A. Colorectal cancer statistics, 2020. *CA: A Cancer Journal for Clinicians* **2020**, *70*, 145.

(12) Gao, D.; Zhang, X.; Liu, B.; Meng, D.; Fang, K.; Guo, Z.; Li, L. Screening circular RNA related to chemotherapeutic resistance in breast cancer. *Epigenomics* **2017**, *9*, 1175−1188.

(13) Hussen, B. M.; Azimi, T.; Hidayat, H. J.; Taheri, M.; Ghafouri-Fard, S. NF-KappaB interacting LncRNA: review of its roles in neoplastic and non-neoplastic conditions. *Biomedicine & Pharmacotherapy* **2021**, *139*, 111604.

(14) Pan, J.-A.; Tang, Y.; Yu, J.-Y.; Zhang, H.; Zhang, J.-F.; Wang, C.-Q.; Gu, J. miR-146a attenuates apoptosis and modulates autophagy by targeting TAF9b/P53 pathway in doxorubicin-induced cardiotoxicity. *Cell Death Disease* **2019**, *10*, 668.

(15) Zhao, L.; Qi, Y.; Xu, L.; Tao, X.; Han, X.; Yin, L.; Peng, J. MicroRNA-140-5p aggravates doxorubicin-induced cardiotoxicity by promoting myocardial oxidative stress via targeting Nrf2 and Sirt2. *Redox biology* **2018**, *15*, 284−296.

(16) Shang, J.; Chen, W.-M.; Liu, S.; Wang, Z.-H.; Wei, T.-N.; Chen, Z.-Z.; Wu, W.-B. CircPAN3 contributes to drug resistance in acute myeloid leukemia through regulation of autophagy. *Leukemia Research* **2019**, *85*, 106198.

(17) Li, Y.; Wang, R.; Zhang, S.; Xu, H.; Deng, L. LRGCPND: Predicting Associations between ncRNA and Drug Resistance via Linear Residual Graph Convolution. *International Journal of Molecular Sciences* **2021**, *22*, 10508.

(18) Dayun, L.; Junyi, L.; Yi, L.; Qihua, H.; Deng, L. MGATMDA: Predicting microbe-disease associations via multi-component graph attention network. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **2021**, 1−1.

(19) Ji, C.; Liu, Z.; Wang, Y.; Ni, J.; Zheng, C. GATNNCDA: A Method Based on Graph Attention Network and Multi-Layer Neural Network for Predicting circRNA-Disease Associations. *International Journal of Molecular Sciences* **2021**, *22*, 8505.

(20) Deng, L.; Huang, Y.; Liu, X.; Liu, H. Graph2MDA: a multi-modal variational graph embedding model for predicting microbe−drug associations. *Bioinformatics* **2022**, *38*, 1118−1125.

(21) Zhao, X.; Zhao, X.; Yin, M. Heterogeneous graph attention network based on meta-paths for lncRNA−disease association prediction. *Briefings in Bioinformatics* **2022**, *23*, bbab407.

(22) Fan, Y.; Chen, M.; Pan, X. GCRFLDA: scoring lncRNA-disease associations using graph convolution matrix completion with conditional random field. *Briefings in Bioinformatics* **2022**, *23*, bbab361.

(23) Wang, L.; You, Z.-H.; Li, Y.-M.; Zheng, K.; Huang, Y.-A. GCNCDA: a new method for predicting circRNA-disease associations based on graph convolutional network algorithm. *PLOS Computational Biology* **2020**, *16*, No. e1007568.

(24) Lan, W.; Dong, Y.; Chen, Q.; Liu, J.; Wang, J.; Chen, Y.-P. P.; Pan, S. IGNSCDA: predicting CircRNA-disease associations based on improved graph convolutional network and negative sampling. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **2021**, 3111607.

(25) Li, H.-Y.; Chen, H.-Y.; Wang, L.; Song, S.-J.; You, Z.-H.; Yan, X.; Yu, J.-Q. A structural deep network embedding model for

predicting associations between miRNA and disease based on molecular association network. *Sci. Rep.* **2021**, *11*, 12640.
(26) Li, L.; Wu, P.; Wang, Z.; Meng, X.; Cai, J.; et al. NoncoRNA: a database of experimentally supported non-coding RNAs and drug targets in cancer. *Journal of Hematology Oncology* **2020**, *13*, 15.
(27) Dai, E.; Yang, F.; Wang, J.; Zhou, X.; Song, Q.; An, W.; Wang, L.; Jiang, W. ncDR: a comprehensive resource of non-coding RNAs involved in drug resistance. *Bioinformatics* **2017**, *33*, 4010−4011.
(28) Li, X.; Zhang, H.; Wang, R.; Nie, F. Multiview Clustering: A Scalable and Parameter-Free Bipartite Graph Fusion Method. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2022**, *44*, 330−344.
(29) Hanley, J. A.; McNeil, B. J. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* **1982**, *143*, 29−36.
(30) Liu, D.; Huang, Y.; Nie, W.; Zhang, J.; Deng, L. SMALF: miRNA-disease associations prediction based on stacked autoencoder and XGBoost. *BMC bioinformatics* **2021**, *22*, 219.
(31) Zeng, M.; Lu, C.; Zhang, F.; Li, Y.; Wu, F.-X.; Li, Y.; Li, M. SDLDA: lncRNA-disease association prediction based on singular value decomposition and deep learning. *Methods* **2020**, *179*, 73−80.
(32) Lu, C.; Zeng, M.; Zhang, F.; Wu, F.-X.; Li, M.; Wang, J. Deep matrix factorization improves prediction of human circRNA-disease associations. *IEEE Journal of Biomedical and Health Informatics* **2021**, *25*, 891−899.
(33) Liu, Y.; Wang, S.-L.; Zhang, J.-F.; Zhang, W.; Zhou, S.; Li, W. DMFMDA: Prediction of microbe-disease associations based on deep matrix factorization using Bayesian Personalized Ranking. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **2021**, *18*, 1763.
(34) Chen, X.; Huang, Y.-A.; You, Z.-H.; Yan, G.-Y.; Wang, X.-S. A novel approach based on KATZ measure to predict associations of human microbiota with non-infectious diseases. *Bioinformatics* **2017**, *33*, 733−739.
(35) Luo, J.; Long, Y. NTSHMDA: prediction of human microbe-disease association based on random walk by integrating network topological similarity. *IEEE/ACM transactions on computational biology and bioinformatics* **2018**, *17*, 1341−1351.
(36) Deepthi, K.; Jereesh, A. Inferring potential CircRNA−disease associations via deep autoencoder-based classification. *Molecular Diagnosis & Therapy* **2021**, *25*, 87−97.
(37) Peng, L.-H.; Yin, J.; Zhou, L.; Liu, M.-X.; Zhao, Y. Human microbe-disease association prediction based on adaptive boosting. *Frontiers in microbiology* **2018**, *9*, 2440.
(38) He, X.; Deng, K.; Wang, X.; Li, Y.; Zhang, Y.; Wang, M. LightGCN: Simplifying and Powering Graph Convolution Network for Recommendation. *arXiV* 2020, https://arxiv.org/abs/2002.02126.
(39) Wang, X.; He, X.; Wang, M.; Feng, F.; Chua, T. S. Neural Graph Collaborative Filtering. *SIGIR'19: Proc. 42nd International ACM SIGIR Conference* **2019**, 165.
(40) Wattenberg, M.; Viégas, F.; Johnson, I. How to use t-SNE effectively. *Distill* **2016**, *1*, No. e2.
(41) Karachi, A.; Dastmalchi, F.; Mitchell, D. A.; Rahman, M. Temozolomide for immunomodulation in the treatment of glioblastoma. *Neuro-oncology* **2018**, *20*, 1566−1572.
(42) Cameron, D.; Gabra, H.; Leonard, R. Continuous 5-fluorouracil in the treatment of breast cancer. *British journal of cancer* **1994**, *70*, 120−124.
(43) Wirsching, H.-G.; Weller, M. Glioblastoma. *Malignant Brain Tumors* **2017**, 265−288.
(44) Vodenkova, S.; Buchler, T.; Cervena, K.; Veskrnova, V.; Vodicka, P.; Vymetalkova, V. 5-fluorouracil and other fluoropyrimidines in colorectal cancer: Past, present and future. *Pharmacology & therapeutics* **2020**, *206*, 107447.

https://doi.org/10.1021/acs.jcim.2c00367
J. Chem. Inf. Model. 2022, 62, 3676−3684