Briefings in Bioinformatics, 22(6), 2021, 1–19

https://doi.org/10.1093/bib/bbab165
Problem Solving Protocol

MDA-GCNFTG: identifying miRNA-disease associations based on graph convolutional networks via graph sampling through the feature and topology graph

Yanyi Chu, Xuhong Wang, Qiuying Dai, Yanjing Wang, Qiankun Wang, Shaoliang Peng, Xiaoyong Wei, Jingfei Qiu, Dennis Russell Salahub, Yi Xiong and Dong-Qing Wei

Corresponding authors: Yi Xiong and Dong-Qing Wei, State Key Laboratory of Microbial Metabolism, Shanghai-Islamabad-Belgrade Joint Innovation Center on Antibacterial Resistances, Joint International Research Laboratory of Metabolic & Developmental Sciences and School of Life Sciences and Biotechnology, Shanghai Jiao Tong University, Shanghai 200030, P.R. China; Peng Cheng Laboratory, Vanke Cloud City Phase I Building 8, Xili Street, Nanshan District, Shenzhen, Guangdong, 518055, P.R China. Tel.: +86 21-34204573; E-mail: xiongyi@sjtu.edu.cn, dqwei@sjtu.edu.cn

Abstract

Accurate identification of the miRNA-disease associations (MDAs) helps to understand the etiology and mechanisms of various diseases. However, the experimental methods are costly and time-consuming. Thus, it is urgent to develop computational methods towards the prediction of MDAs. Based on the graph theory, the MDA prediction is regarded as a node classification task in the present study. To solve this task, we propose a novel method MDA-GCNFTG, which predicts MDAs based on Graph Convolutional Networks (GCNs) via graph sampling through the Feature and Topology Graph to improve the training efficiency and accuracy. This method models both the potential connections of feature space and the structural relationships of MDA data. The nodes of the graphs are represented by the disease semantic similarity, miRNA functional similarity and Gaussian interaction profile kernel similarity. Moreover, we considered six tasks simultaneously on the MDA prediction problem at the first time, which ensure that under both balanced and unbalanced sample

Yanyi Chu is a PhD candidate at the School of Life Sciences and Biotechnology, Shanghai Jiao Tong University. She works on drug discovery through machine learning methods.
Xuhong Wang is a PhD candidate at the School of Electronic, Information and Electrical Engineering (SEIEE), Shanghai Jiao Tong University. His research interests lie in dynamic graph neural networks, graph-based anomaly detection and (graph) stream data processing.
Qiuying Dai is a PhD candidate at the School of Life Sciences and Biotechnology, Shanghai Jiao Tong University. Her major research interests include the prediction of miRNA-disease associations and computer-aided drug design.
Yanjing Wang is a postdoctoral scholar at the School of Life Sciences and Biotechnology, Shanghai Jiao Tong University. She works on drug discovery through machine learning methods and molecular dynamics simulations.
Qiankun Wang is a PhD candidate at the School of Life Sciences and Biotechnology, Shanghai Jiao Tong University. He has expertise in computer-aided drug design and machine learning.
Shaoliang Peng is a Professor at the College of Computer Science and Electronic Engineering, Hunan University. His research interests are high-performance computing, bioinformatics, big data, precision medicine, health informatics and computer-aided drug design.
Xiaoyong Wei is a Professor at the Pengcheng Laboratory. His research interests include multimedia retrieval, data mining and machine learning.
Jingfei Qiu is a professor at the Pengcheng Laboratory. His research interests are large data analysis, computer vision, natural language processing and machine learning.
Dennis Russell Salahub is a full Professor at the Department of Chemistry, University of Calgary, Fellow Royal Society of Canada and Fellow of the American Association for the Advancement of Science.
Yi Xiong is an Associate Professor at the School of Life Sciences and Biotechnology, Shanghai Jiao Tong University. His main research interests focus on machine learning algorithms, and their applications in protein sequence–structure–function relationships and biomedicine.
Dong-Qing Wei, professor of bioinformatics, made many groundbreaking contributions to the development of bioinformatics techniques and their interdisciplinary applications to systems of ever-increasing complexity and published more than 400 papers, with 9000 citations and an H factor of 55.
Submitted: 10 February 2021; **Received (in revised form):** 2 April 2021

distribution, MDA-GCNFTG can predict not only new MDAs but also new diseases without known related miRNAs and new miRNAs without known related diseases. The results of 5-fold cross-validation show that the MDA-GCNFTG method has achieved satisfactory performance on all six tasks and is significantly superior to the classic machine learning methods and the state-of-the-art MDA prediction methods. Moreover, the effectiveness of GCNs via the graph sampling strategy and the feature and topology graph in MDA-GCNFTG has also been demonstrated. More importantly, case studies for two diseases and three miRNAs are conducted and achieved satisfactory performance.

Key words: miRNA-disease associations; the feature and topology graph; graph convolutional network; graph sampling

Introduction

MiRNA is a type of endogenous regulatory noncoding RNA discovered in 1993, and its length is about 22 nucleotides [1, 2]. It plays a vital role in a variety of biological processes by targeting specific mRNA and regulating gene expression [3–7], including immune reaction [8], cell cycle regulation [9], tumor invasion [10], etc. In addition, it is proven that miRNAs regulate more than one-third of genes [11], so the dysregulation of miRNAs can lead to cell behavior disorders [12]. Furthermore, many studies have proved that miRNAs are highly correlated with the development of complex human diseases [13–16], especially cancers [17], such as breast cancer [18, 19], lung cancer [20, 21], lymphoma [22] and so on. Therefore, miRNAs may be used as potential biomarkers in the diagnosis of diseases [19, 23, 24]. Thus, identifying the associations between miRNAs and diseases can not only improve the understanding of disease mechanisms but also assist in disease prevention, diagnosis and treatment [25, 26]. Although experimental methods to identify the miRNA-disease association (MDA) have high accuracy, they are very time-consuming and costly. Therefore, the development of computational methods to identify MDAs is necessary and becomes an auxiliary step for experimental methods [27].

Network science is established as a backbone for exploring complex biological systems (i.e. molecular interaction networks). They are graphs composed of biomolecules as nodes and interconnections between biomolecules as edges, such as MDAs studied in this work. A large number of studies have shown that biomolecules do not perform their biological functions alone but express their functions through the interaction with other biomolecules to form a hierarchical community structure [28]. Further, the disease should be described as a 'network disease', because it is rarely caused by a single gene abnormality, but by disturbance or malfunction of the complex biological network of tissues and organ systems [29]. Therefore, the inference of association between biomolecules should consider the network topology. Graph neural networks (GNNs) [30] represent a significant stride to operate directly on network/graph-structured data, and a promising method to address the above problem. GNN is essentially a neighborhood node aggregation scheme, where each node aggregates feature information of its directed neighbors to compute its new feature vector. After multiple iterations of information aggregation, the computed node embedding will capture the structural information among the neighbors of the node. GNNs are being widely used in various real-world tasks and have been achieved satisfactory performance in bioinformatics applications, such as drug–target interactions or affinity predictions [31–36], drug–drug interaction predictions [37–40], disease–gene association identification [41–44], etc.

Graph convolutional network (GCN) [45] is an important branch of GNN and has made great progress in recent years. However, traditional GCN methods usually require full graph training. In MDA or other bioinformatics tasks, the number of related entities (such as drugs, proteins, miRNAs, etc.) is very large. Thus, blindly performing full graph training will cause huge computational complexity due to the 'neighbor explosion' phenomenon and may cause insufficient memory due to too many computing resources being required. Then, most of the work [46–50] is exploring how to reduce training costs by sampling the nodes of each layer of GCNs. However, these methods still face challenges in accuracy, scalability and training complexity [51, 52]. Thus, the subgraph-based methods [51, 53, 54] are designed to suit large graphs and deep networks. Inspired by their ideas, this study samples subgraphs of the original graph and runs the full GCN model on the subgraph for each minibatch. To ensure that these subgraphs retain most of the original edge while still presenting a meaningful topology, we performed an edge-based sampling strategy and added normalization and variance reduction technology.

On the other hand, most existing MDA prediction methods are trained and tested on balanced data, such as [55–58]. They regard the known MDAs as the positive samples, the unknown MDAs as the negative samples and then sample the same number of negative samples as that of the positive samples so that the ratio of positive to negative samples is 1:1. It is worth noting that the distribution of these balanced data does not conform to the natural distribution of MDAs. Although many methods have achieved good performance on these balanced data, it does not mean their high performance on real MDA prediction tasks, because the test set is incomplete. Therefore, it is necessary to consider natural unbalanced data, although the imbalance problem is still a major challenge for machine learning methods [59]. On the other hand, the existing methods only consider the prediction of new miRNA-disease pairs (MDPs) when training and testing, that is, the task pairs (i.e. Tp) in this study. Although most of the current methods have carried out case studies on certain diseases, it is still not enough to account for the overall predictive performance of new miRNAs and diseases that did not appear in the training set. Therefore, this study considered the above two types of viewpoints at the same time and proposed six experimental tasks on the MDA prediction problem for the first time, namely, to predict new MDPs (Tp), predict new miRNAs (Tm) and predict new diseases (Td) on balanced and unbalanced data, respectively. It is worth noting that the positive sample corresponding to the new object in the above tasks is only in the test set, not in the training set.

This research proposes a novel MDA-GCNFTG method (Figure 1) for MDA prediction and implements it on six different prediction tasks. The method is mainly composed of two parts. First, we define the feature and topology graph that fully explore node (i.e. miRNAs and diseases) feature, network topology (i.e. MDAs or miRNA-disease links) and their combination through the k-nearest neighbors (k-NN) algorithm [60, 61] to introduce

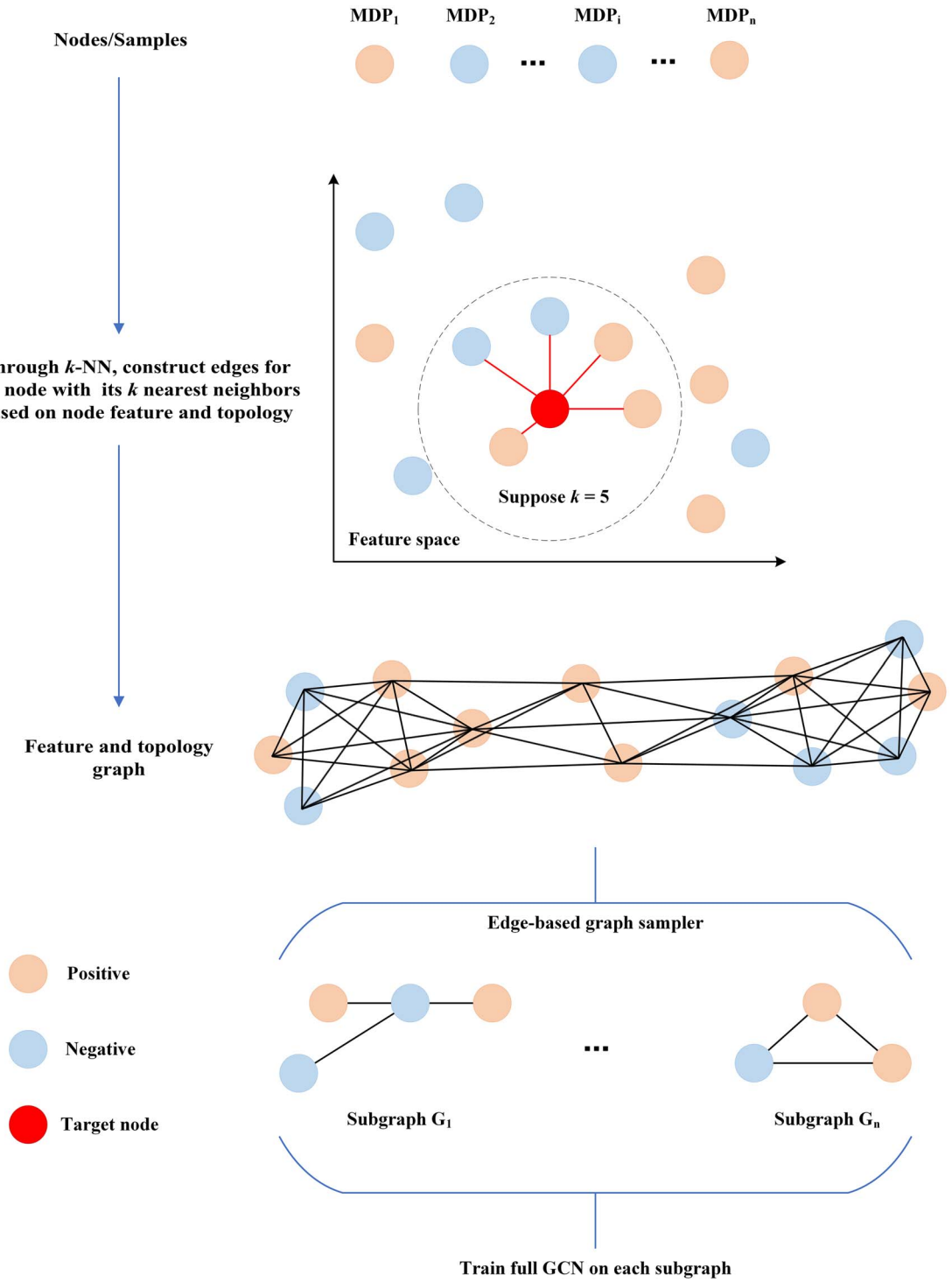

Figure 1. The workflow of the proposed MDA-GCNFTG method, where the MDP represents the miRNA-disease pair, GCN represents GCN.

the most helpful and deepest relevant information for the MDA prediction task. For this graph, the node is the MDP, the node label represents whether the MDP is MDA or not and the edge is constructed between the node and its k nearest neighbors based on the node information. It is worth noting that regarding the use of MDP as the node, we considered two reasons: (i) Based on the assumption that similar miRNAs are more likely to be related to similar diseases and vice versa [62], further, similar MDPs tend to have similar associations (i.e. labels). Implementing the GCN algorithm on this graph will make the similar nodes (i.e. similar miRNA-disease pairs) to be clustered. (ii) Compared with heterogeneous graphs, homogeneous graphs are easier to learn. Then, a novel GCN algorithm based on graph sampling is implemented on the feature and topology graph. The experimental results show that the proposed MDA-GCNFTG method has achieved satisfactory results in all six tasks

and is superior to several classic machine learning algorithms and state-of-the-art methods on the MDA prediction problem. Moreover, this research also demonstrates the effectiveness of k-NN and the novel GCN algorithm in this method. More importantly, we conducted two types of case studies for miRNAs and diseases, respectively. The results demonstrate the satisfactory performance and prove the effectiveness of the proposed MDA-GCNFTG method.

Related work

In recent years, a large number of computational methods [63] have been developed for MDA prediction problems and can be divided into four categories [27], including score function–based, multiple biological information–based, complex network algorithm–based, and machine learning–based methods. The score function–based methods [62, 64–66] define the score function based on the probability distribution or statistical analysis of training data to measure the degree of the potential MDA. The multiple biological information–based methods [67–99] consider the bioinformatics knowledge related to miRNAs and diseases, which may also include entities of circRNA, mRNA, lncRNA, drug, protein, microbe, etc. The heterogeneous graph constructed through the above entities and the relationship among them can provide valuable information for constructing the relationship between miRNAs and diseases. The complex network algorithm–based methods [100–191] predict potential MDAs mainly based on various disease and miRNA similarity networks from different perspectives. The machine learning–based methods [55, 192–228] are the important branch in the field of MDA prediction. They use machine learning algorithms to extract effective feature representations and solve the specific optimization problem to obtain reliable MDA prediction. It is worth noting that the above four types of methods are not without intersection. For example, the GraRep method [229] adopts the ideas of multiple biological information–based, complex network algorithm–based, and machine learning–based methods at the same time. It establishes a heterogeneous graph network containing miRNA, disease, drug, protein, lncRNA and the associations among them. In the construction of embedding representations, the similarity information of disease was also considered. Finally, the random forest (RF) algorithm that belongs to the machine learning algorithm was used to predict the potential MDA.

Some studies have tried to apply GCNs to the MDA prediction problem, and they can be categorized into the following four categories. (i) Pairwise GCNs methods [57, 230, 231], which use two GCNs to extract miRNAs and diseases embedding, and then predict MDAs. This type of method does not consider the connection among MDPs. (ii) The link prediction method of bipartite graph [56], which uses miRNAs and diseases as nodes, MDAs as edges and GCNs to predict potential MDAs. It regards negative samples as a kind of edges to participate in node update, which causes the oversmoothing problem caused by too many false neighbors. (iii) The GCN method based on the fully connected graph [58]. Since the graph is too dense, after the nodes are updated, the embedding of each node tends to be unified, which causes the oversmoothing problem. (iv) Pan *et al*. proposed studies [232, 233] that use the multilabel GCN to infer disease-associated miRNAs in a semisupervised manner. However, these two methods only realize a part of the MDA prediction, without considering the task for prediction of its associated miRNAs for a given disease.

Table 1. Summary of the corresponding miRNA-disease associations' information in the test data of three experimental settings

Experimental settings	Diseases	miRNAs	Associations
Tp	Known	Known	New
Td	New	Known	New
Tm	Known	New	New

Table 2. Summary of the samples on the balanced and unbalanced data

Data	Known associations	Unknown associations
Balanced	5430	5430
Unbalanced	5430	184 155

Materials and methods

Data set

In this study, the human miRNA disease database (HMDD) v2.0 [234] is adopted as the benchmark data set. There are 5430 experimentally verified MDAs consisting of 495 miRNAs and 383 diseases. In the MDA prediction problem, known MDAs are considered as positive samples, and the negative data contain all unknown or nonexisting MDAs.

The HMDD v2.0 used in this study was released in 2014. Some recent studies [55, 71, 73, 172, 176, 209] have used the updated data set of HMDD v3.0 [235] released in 2019. Therefore, we also used the larger and new version HMDD v3.0 to train and test our method. The statistical information of HMDD v3.0 is shown in Supplementary Table S1, see Supplementary Data available online at http://bib.oxfordjournals.org/.

Experiment settings and tasks

This study evaluates the MDA prediction problem through three experiment settings: (i) task pairs (Tp), which predicts new MDPs; (ii) task diseases (Td), which predicts new diseases and (iii) task miRNAs (Tm), which predicts new miRNAs. The label of the new object in the corresponding task is missed in the training set, but it exists in the test set to predict and evaluate the model performance (as shown in Table 1).

On the other hand, we evaluate the above three experimental settings on balanced and unbalanced data. For the unbalanced task, we consider the entire space of MDAs in these three tasks to simulate more practically, that is, use all negative data as negative samples to participate in training and test. Therefore, the number of positive samples is much lower than the number of negative samples, resulting in an imbalance of data. Moreover, we also use the balanced data to follow the previous work, that is, the same amount of data as the positive sample is sampled in the negative data as the negative sample before training and test. The details are as shown in Table 2.

Finally, this study performed six tasks to cover most cases predicted by MDAs.

Node feature construction

This study adopted an integrated feature based on the diseases semantic similarity, miRNAs functional similarity and Gaussian interaction profile (GIP) kernel similarities. The feature generation process is shown in Figure 2.

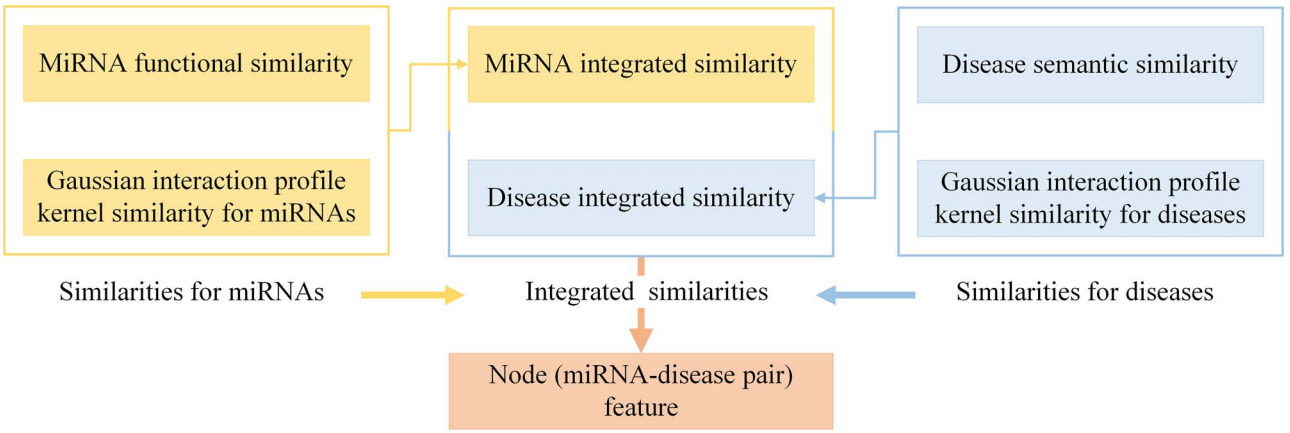

Figure 2. The feature generation process of the proposed MDA-GCNFTG method.

The miRNA functional similarity matrix

The miRNA functional similarity matrix (MFSM) is built based on the assumption that miRNAs with similar functions are more likely to be associated with diseases with similar phenotypes, and vice versa [236]. The similarity information can be obtained from https://www.cuilab.cn/files/images/cuilab/misim.zip.

The disease semantic similarity matrix

The disease semantic similarity can be calculated based on the medical subject headings descriptors [195], which can be obtained from https://www.ncbi.nlm.nih.gov/. Many studies [56, 62, 236] used the directed acyclic graph (DAG) to generate disease semantic similarity matrix (DSSM), where the DAG describes the relationship of different diseases.

There are two different DSSMs defined from two considerations. The $DSSM_1$ is generated based on the assumption that if two diseases share a larger part of their DAGs, they can be considered more similar. The $DSSM_2$ further considered that if the disease appears in more (or less) DAGs, it may be more common (or specific). Therefore, in the same layer of DAG, the semantic contribution value of diseases should be different. These two DSSMs are obtained from GAMEDA [56].

In order to obtain a more reasonable DSSM, we perform element-wise averaging on the above two DSSMs to synthesize the final DSSM.

The GIP kernel similarity

Based on the assumption that similar miRNAs are more likely to be related to similar diseases [62], the GIP kernel similarity matrix for miRNAs (MGSM) and diseases (DGSM) can be calculated.

Take building DGSM as an example. First, build the miRNA interaction profile for diseases; the column i represents the miRNA interaction profile y_{d_i} of the disease d_i. It is a binary vector, and each element represents the association between the disease and the corresponding miRNA. If there is an association, the element value is 1; otherwise, it is 0. Next, calculate the GIP similarity between disease d_i and d_j according to the two corresponding columns of the miRNA interaction profile:

$$\mathrm{DGSM}(d_i, d_j) = \exp\left(-\left(\frac{n_d}{\sum_{m=1}^{n_d}\left\|y_{d_m}\right\|^2}\right)\left\|y_{d_i} - y_{d_j}\right\|^2\right) \quad (1)$$

where n_d is the number of diseases. Moreover, due to normalization, this kernel is independent of the size of the data set.

The MGSM is calculated in the similar way as DGSM.

Integrated similarity as node feature

Considering that there are many sparse values in the MFSM and DSSM obtained above, we fuse the GIP kernel similarity MGSM and DGSM to fill the zero-values, respectively. Then, the integrated miRNA and disease similarity matrix (that are IM and ID) are obtained. The integrated equations [62] are

$$\mathrm{IM}(m_i, m_j) = \begin{cases} \mathrm{MFSM}(m_i, m_j) & \text{if MFSM}(m_i, m_j) \text{ not equal to 0} \\ \mathrm{MGSM}(m_i, m_j) & \text{otherwise} \end{cases} \quad (2)$$

$$\mathrm{ID}(d_i, d_j) = \begin{cases} \mathrm{DSSM}(d_i, d_j) & \text{if DSSM}(d_i, d_j) \text{ not equal to 0} \\ \mathrm{DGSM}(d_i, d_j) & \text{otherwise} \end{cases}. \quad (3)$$

Then, the IM and ID are spliced as the node (i.e. MDP) feature of the feature and topology graph for the proposed MDA-GCNFTG method.

Methods

There are two crucial steps in the MDA-GCNFTG method (Figure 1): (i) construct the feature and topology graph through integrated similarity and k-NN algorithm and (ii) predict the MDA by a novel GCN algorithm via graph sampling.

Preliminaries

Define a graph G (V, E, X), where V is the node set, E is the edge set and X is the node feature matrix. This graph describes the relationship among nodes with attributes. Applying the GCN algorithm to the graph data to predict the category of each node is the node classification task. In order to classify nodes, GCNs use the feature of the node itself, as well as the information of neighboring nodes and edges for message passing, which can be performed multiple times to aggregate information from a wider range of neighbor nodes to update the node feature.

GCN is a neural network layer, and its propagation mode from layer l to layer $l+1$ is

$$H^{(l+1)} = \sigma\left(\tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}H^{(l)}W^{(l)}\right), \tilde{A} = A + I, \tilde{D}_{ii} = \sum_j \tilde{A}_{ij}, \quad (4)$$

where A is the adjacency matrix, I is the identity matrix, $H^{(l)}$ is the feature of the lth layer,$W^{(l)}$ is the weight of the lth layer and σ is a nonlinear activation function. For the input layer, H is X.

For the node classification problem, suppose we construct two-layer GCNs, and the activation function adopts ReLU and softmax, respectively, then the overall forward propagation formula is

$$\hat{Y} = \text{softmax}\left(\hat{A}\text{ReLU}\left(\hat{A}XW^{(0)}\right)W^{(1)}\right), \hat{A} = \tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}. \quad (5)$$

Finally, the cross-entropy loss function is calculated for all labeled nodes.

Construct the feature and topology graph through the k-NN algorithm

There is a study that revealed [237] that the ability of GCNs to integrate network topology and node features to extract the most relevant information for the task is not ideal, which may seriously hinder the performance of classification tasks. Furthermore, similarities between the feature and the information inferred from the topological structure are complementary to each other, and fusing them can get deeper related information for classification tasks [237]. Moreover, the correlation between graph data and tasks is often complicated and unknowable, so adaptive ability is also important in practical applications.

Inspired by the above study, this study proposes an adaptive graph-building method, which can adaptively propagate node features and topological information to the feature space. In order to fully capture the structural information in the feature space, we generate feature and topology graph through the k-NN algorithm based on the node feature and the topological relationship between miRNAs and diseases. To realize this point, the MDP is used as the node in the graph, the node feature is the integrated similarity of the miRNA and the disease, and the label of each node represents whether there is an association between the miRNA and the disease. Compared with existing GCN-based MDA prediction methods, this graph-building strategy not only takes associations among MDPs into account but also realizes the effective fusion of heterogeneity. Finally, a homogeneous graph is generated to do the node classification task for the MDA prediction problem.

The procedure to generate the feature and topology graph is to fit a k-NN classifier, predict the label for each node and use the k nearest correctly classified nodes as neighbors of the node. Obviously, the result of the k-NN algorithm largely depends on the choice of k. Thus, we tune the k parameter (i.e. at the value of 1, 3, 5, 7, 10 or 15) to study the influence of the number of neighbors on the MDA prediction. On the other hand, this k-NN algorithm is performed on all data. To guarantee that the test data are not leaked in the training phase, we set the label of nodes that belong to test data to 0.

GCNs based on graph sampling and normalization

In this study, a novel GCN algorithm is applied to MDA prediction tasks, and the overall training algorithm is illustrated in Algorithm 1.

This algorithm is different from the traditional GCN algorithm in minibatch construction; it is based on graph sampling. The idea is to sample multiple subgraphs of the training graph first and then construct the complete GCN on each subgraph. In this way, when propagating in the GCN layer, accurate node embedding can be obtained from the subgraph, and the sampled nodes can support each other without collecting information from outside the batch. Naturally, this algorithm solves the dilemma of the neighbor explosion, which is usually encountered by traditional GCN algorithms. In order to ensure the accuracy of training, a suitable graph sampler is needed. First of all, we consider that nodes that have a great influence on each other should be sampled in the same subgraph, so this study uses a topology-based edge sampler. But this influence-driven sampling idea will introduce bias. In order to eliminate this bias, this algorithm introduces the sampling probability of nodes and edges to carry out normalization when aggregating node information and calculating the minibatch loss.

When defining an edge sampler, the main point is that edges with a nonnegligible probability should be sampled. On the other hand, it also considers that the variance of the aggregation of the node in full GCNs should be reduced. According to this, the optimal edge sampling probability is defined (see the 7th row of Algorithm 1). The formula shows that if two connected nodes u and v have very few neighbors (that is, they are very influential to each other), then the edge probability $p_{(u,v)} = p_{(v,u)}$ will be high.

Performance evaluation

In order to facilitate the comparison with other methods, we followed previous studies [56, 62, 152, 156, 195, 197, 229] and performed 5-fold cross-validation (CV) for each task. We also carried out global and local leave-one-out CV (Details are shown in Supplementary Section 2, see Supplementary Data available online at http://bib.oxfordjournals.org/).

For each fold of each task, the following metrics are calculated:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (6)$$

$$\text{Precision} = \frac{TP}{TP + FP} \quad (7)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (8)$$

$$\text{F1} - \text{score} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}, \quad (9)$$

where TP is true positive, FP is false positive, FN is false negative and TN is true negative in the predicting results.

Moreover, the area under the precision–recall curve (AUPR) and the receiver operating characteristic curve (AUC) are also calculated.

It is worth noting that due to the essential difference between balanced and unbalanced data, the importance of different metrics is also different when performing performance evaluation. In the balanced task, all the above six metrics are important. It is worth noting that recall and precision are usually a pair of contradictory performance metrics, so F1-score is often used to characterize their comprehensive performance. Therefore, in the following discussion, the average of the performance evaluation metrics is calculated from accuracy, F1-score, AUC and AUPR for balanced tasks. However, in unbalanced tasks, accuracy is less meaningful, and AUPR can provide better performance estimates than AUC because it will punish false positives more severely. Therefore, the average performance evaluation metric in the following discussion is calculated from F1-score and AUPR.

Hypothesis test

When comparing multiple algorithms on a set of tasks, Demšar [238] recommends the Friedman rank test [239, 240], which uses rank to realize a nonparametric test to validate whether there

Algorithm 1. GCN based on graph sampling and normalization

Input: Training graph $G(V, E, X)$; Labels Y; The number of subgraphs N in pre-processing; Edge budget m
Output: GCN model with trained weights

function Edge sampler (G, m, P)
$E_s \leftarrow m$ edges randomly sampled from E according to P
$V_s \leftarrow$ Set of nodes that are end-points of edges in E_s
$G_s\ (V_s, E_s) \leftarrow$ Node induced subgraph of G from V_s
end function
function Pre-processing (G, N)
The probability of an edge (u, v) being sampled in a subgraph $P(e_{u,v}) = (\frac{1}{\deg(u)} + \frac{1}{\deg(v)}) / \sum_{(u',v')\in E}(\frac{1}{\deg(u')} + \frac{1}{\deg(v')})$
$G_{s,n}\ (V_s, E_s), n = 1, \ldots, N \leftarrow$ Repeatedly run the Edge sampler N times to obtain N subgraphs of G
$C_v \leftarrow$ for each node $v \in V$, count the number of times the node appears in the N subgraphs
$C_{u,v} \leftarrow$ for each edge $(u, v) \in E$, count the number of times the edge appears in the N subgraphs
Normalization coefficients $\alpha_{u,v} = \frac{C_{u,v}}{C_v} = \frac{C_{v,u}}{C_v}$ and $\lambda_v = \frac{C_v}{N}$
end function
Run Pre-processing to obtain N G_s, which can be reused in training, and compute the edge probability P and normalization coefficients α, λ
for each minibatch **do**
$G_s\ (V_s, E_s) \leftarrow$ According to Edge sampler, sampled subgraph of G
Construct GCN on G_s
$\{\hat{y_v} \mid v \in V_s\} \leftarrow$ Forward propagation of $\{x_v \mid v \in V_s\}$, normalized by α
Update weights through backward propagation from λ-normalized loss $\text{Loss}_{\text{batch}} = \sum_{v\in G_s} \frac{\text{Loss}_v(\hat{y_v}, y_v)}{\lambda_v}$
end for

are significant differences between multiple overall distributions. In this study, the null hypothesis is that there are no differences among different methods. The statistical result of the hypothesis test (that is, the P-value of the Friedman test) is used to determine whether to reject the null hypothesis or not based on the significance level α. If the null hypothesis is rejected, that is, the difference between at least two methods is statistically significant, we will subsequently compare every two methods in pairs. In the pairwise comparison analysis by Friedman test, we used the Bonferroni-adjusted P-value to take into account the problem of type I error expansion in multiple comparisons, so the accuracy is better than using the original P-value. Finally, this method can indicate whether there is a significant difference between different methods.

Results and discussion

Effect of the k-NN algorithm in the proposed MDA-GCNFTG method

In order to verify the effectiveness of the k-NN algorithm in the MDA-GCNFTG method, we first compared the edge-building methods of 1-NN and self-loop, because 1-NN is similar to self-loop that both of them create only one edge for each node in the graph. The self-loop establishes an edge from a node to the node itself, and each node in the graph has no neighbor nodes, so effective node updates cannot be performed. The self-loop strategy, as the baseline of the edge construction method in MDA-GCNFTG, represents the lowest performance of the proposed MDA-GCNFTG method. 1-NN establishes an edge between each node and one of its neighbors. Although it can perform effective node updates, the degree of nodes in the graph is too low (only 1), resulting in low graph utilization. Thus, the performance of 1-NN is also low in MDA-GCNFTG. The results of the comparison are presented in Figure 3. It can be seen that 1-NN is better than the self-loop method in six tasks, especially in the balanced task. This proves that introducing links among nodes by the k-NN algorithm can improve the performance of MDA tasks.

Then, we clarify the robustness of the model by comparing different numbers of closest neighbors (i.e. k) in the k-NN algorithm. For this purpose, the value of k includes 1, 3, 5, 7, 10 and 15. The results show (Figure 3) that for each task, the prediction performance of different k is approximately equal to each other. It indicates that the proposed MDA-GCNFTG is not very sensitive to k and its robustness to the edge-building step is proven, which will avoid a lot of work in parameter tuning. Moreover, MDA-GCNFTG also has versatility for different tasks, so it can be migrated to other MDA applications. The above point of view has also been confirmed by experiments conducted on HMDD v3.0 (see Supplementary Tables S2–S4, see Supplementary Data available online at http://bib.oxfordjournals.org/).

In addition, we also proved the above viewpoint from a statistical perspective. The data for the Friedman test of each algorithm are all the performance evaluation metrics on different tasks. According to the average rank of each method calculated by the Friedman test, the self-loop method ranks last (that is, the lowest average rank), followed by 1-NN, indicating that these two methods are indeed inferior to other k-NN methods in MDA-GCNFTG. Through the pairwise comparison of the Friedman test, self-loop has significant differences with all the six k-NN methods. Its Bonferroni-adjusted P-value with 1-NN is 0.018, and its Bonferroni-adjusted P-value with the other five k-NN methods is less than 0.001. This result proves that the similar edge-building strategies of self-loop and 1-NN both lead to similar lower performance. In addition, the Bonferroni-adjusted P-values between 1-NN with 5-NN and 7-NN are 0.025 and 0.005, respectively, which indicates the difference between them is also statistically significant.

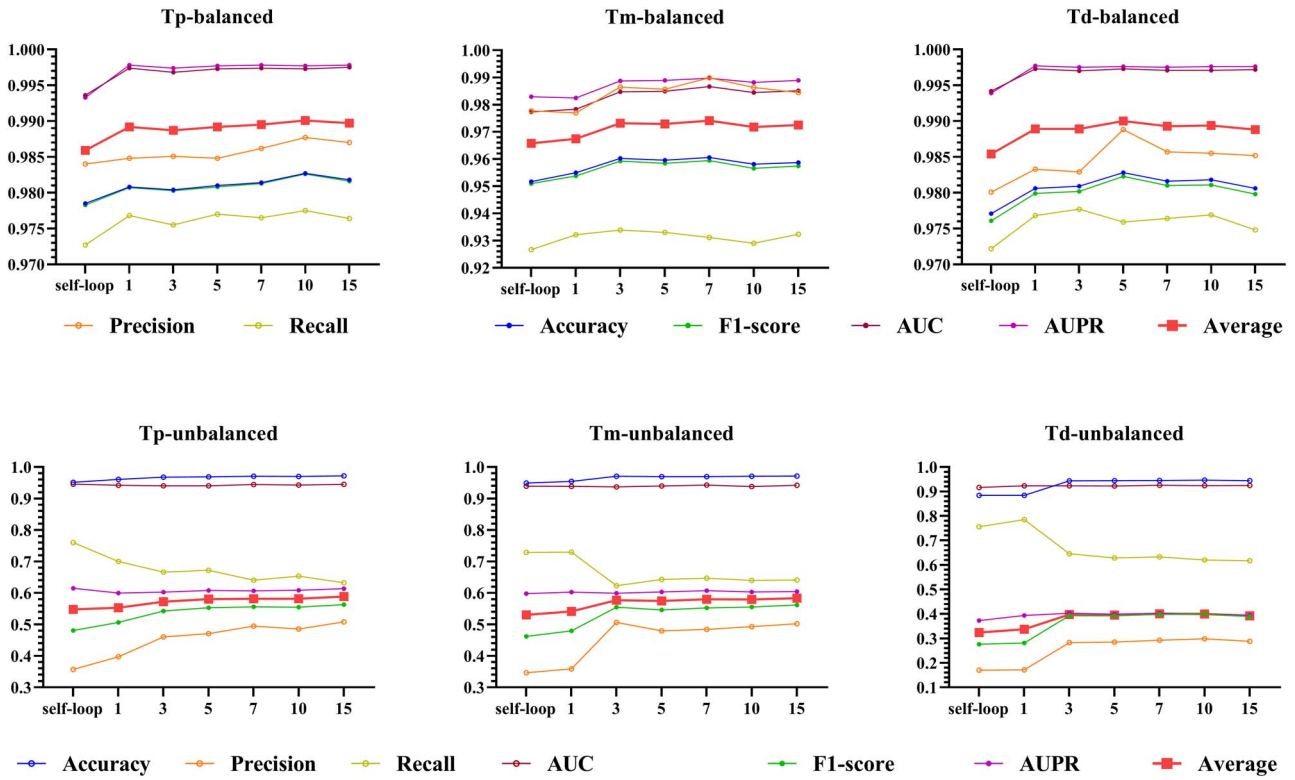

Figure 3. The effect of *k*-NN for MDA-GCNFTG's performance on six tasks, where the abscissa axis represents the self-loop and the number of neighbors. For balanced tasks, the average value is calculated from Accuracy, F1-score, AUC and AUPR. For unbalanced tasks, the average value is calculated from F1-score and AUPR.

The above experiments show that the *k*-NN algorithm can adaptively extract the most relevant information for different tasks and improve the classification performance.

Performance of MDA-GCNFTG in different MDA prediction tasks

Table 3 shows the performance of the proposed MDA-GCNFTG method for six tasks, where each task uses the prediction model obtained after the edge is established from the optimal *k* value. The results show that MDA-GCNFTG has demonstrated its extraordinary predictive ability on balanced tasks, and most of the metrics have reached 0.98. On unbalanced tasks, MDA-GCNFTG does not seem to have high performance, but the discussion in subsequent sections will show its superiority.

On the other hand, this is the first time that the MDA prediction problem has been involved in tasks other than Tp-balanced tasks, and all of them have shown satisfactory performance. It is worth noting that Tm and Td tasks are more difficult than Tp. Because their goal is to predict new miRNAs and new diseases, that is, predict objects that have never appeared in the training set. In addition, we conducted three balanced tasks on HMDD v3.0, and the results can be seen in Supplementary Tables S2–S4, see Supplementary Data available online at http://bib.oxfordjournals.org/. In order to follow the previous research [207, 210, 241], we also performed global and local leave-one-out CV on the traditional Tp-balanced task. The satisfactory results can be seen in Supplementary Section 2, see Supplementary Data available online at http://bib.oxfordjournals.org/.

Effect of the novel GCN algorithm in the proposed MDA-GCNFTG method

In order to prove that the novel GCN algorithm proposed in this study is effective on the MDA prediction task, we compared it with the traditional GCN algorithm. When implementing the GCN algorithm, its experimental conditions are exactly the same as the MDA-GCNFTG method, including 5-fold CV, data partitioning, random seeds, edge or graph constructions, etc.

The results show (Figure 4) that the proposed MDA-GCNFTG method has higher performance than GCN on all six tasks. According to the average performance evaluation metric, its superiority on two difficult tasks (Tp and Tm tasks on unbalanced data) is more significant. This not only illustrates the effectiveness of this novel GCN algorithm on MDA prediction tasks but also proves the superiority of the novel GCN algorithm and the MDA-GCNFTG method proposed in this study compared to the traditional GCN method. The above point of view has also been confirmed by experiments conducted on HMDD v3.0 (see Supplementary Tables S5–S7, see Supplementary Data available online at http://bib.oxfordjournals.org/).

We also compared the time and memory differences between the proposed novel GCN algorithm and the traditional GCN algorithm on Nvidia GeForce RTX 3080 with 10,018 MB memory. All the experimental conditions of this experiment are the same as the above except the epoch is set to one. The results are shown in Table 4. For balanced tasks, the memory consumption of the two methods is very close. But in terms of running speed, MDA-GCNFTG has obvious advantages, especially in Tp and Td tasks, which is twice as fast as GCNs. On unbalanced tasks, GCNs cannot run on Tp and Tm tasks due to insufficient memory, and it runs very slowly on the CPU. MDA-GCNFTG can run on all

Table 3. The 5-fold CV results for six tasks, where the Std and Avg represent standard deviation and average value, respectively

Tasks	Fold	Accuracy	Precision	Recall	F1-score	AUC	AUPR
Tp-balanced	Std	0.0035	0.0058	0.0021	0.0036	0.0005	0.0004
	Avg	0.9827	0.9877	0.9775	0.9826	0.9973	0.9977
Tm-balanced	Std	0.0153	0.0036	0.0289	0.0148	0.0050	0.0036
	Avg	0.9606	0.9898	0.9312	0.9594	0.9866	0.9897
Td-balanced	Std	0.0017	0.0072	0.0063	0.0037	0.0005	0.0006
	Avg	0.9828	0.9888	0.9759	0.9823	0.9973	0.9976
Tp-unbalanced	Std	0.0019	0.0248	0.0220	0.0121	0.0024	0.0098
	Avg	0.9718	0.5080	0.6324	0.5628	0.9448	0.6137
Tm-unbalanced	Std	0.0081	0.0406	0.0411	0.0255	0.0098	0.0173
	Avg	0.9710	0.5024	0.6412	0.5617	0.9421	0.6042
Td-unbalanced	Std	0.0098	0.0429	0.0350	0.0417	0.0105	0.0542
	Avg	0.9451	0.2923	0.6325	0.3981	0.9253	0.4026

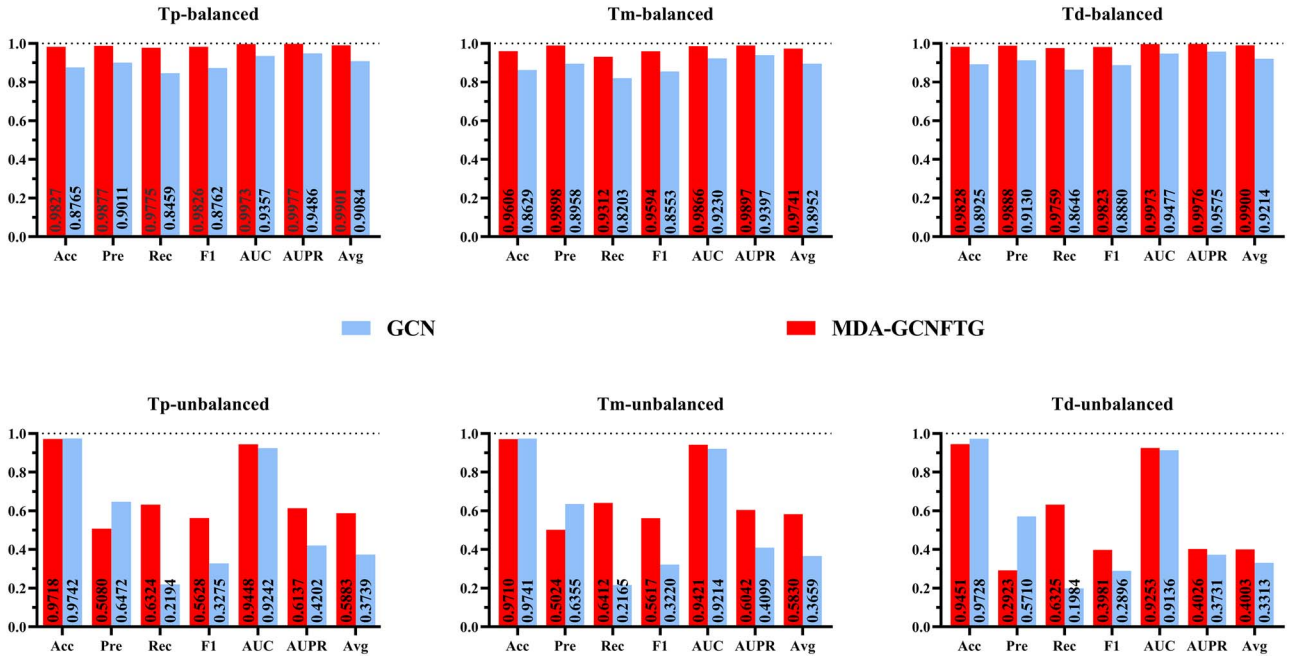

Figure 4. The comparison of the proposed MDA-GCNFTG method with the traditional GCN method, which reflects the effect of the novel GCN for MDA-GCNFTG's performance on six tasks. Acc, Pre, Rec and F1 represent the accuracy, precision, recall and F1-score, respectively. For balanced tasks, the average value is calculated from accuracy, F1-score, AUC and AUPR. For unbalanced tasks, the average value is calculated from F1-score and AUPR.

three unbalanced tasks and only uses about 7000 MB of memory. On the Td task, MDA-GCNFTG not only consumes less memory than the GCN method, but it also runs much faster than the GCN method, i.e. their running times under one epoch are 14 and 65 s, respectively.

Comparisons of MDA-GCNFTG with classic machine learning models

In order to illustrate the superiority of the GCNFTG model proposed in this study in MDA prediction, we also compare it with some classic machine learning algorithms, including deep learning–based deep neural network (DNN), RF, extremely randomized trees (ERTs), decision trees (DTs) and Gaussian naïve Bayes (GNBs). The results of the above models on six tasks are shown in Figure 5.

According to the results, the proposed GCNFTG model is superior to other machine learning models, especially in balanced tasks. For unbalanced tasks, although the recall of MDA-GCNFTG is lower than other models, we should realize that precision and recall are mutually contradictory metrics. Therefore, the F1-score that combines the two is worthy of attention, and it shows the superiority of MDA-GCNFTG compared to other methods. Moreover, the average of F1-score and AUPR proves this again. In order to fully test the superiority of the proposed MDA-GCNFTG method compared to other classic machine learning algorithms, we conducted a Friedman test. Although different performance evaluation metrics have different importance on unbalanced tasks, the test is still performed on all performance evaluation metrics. The results show that the MDA-GCNFTG method has the largest average rank, i.e. ranking first. Moreover, the MDA-GCNFTG method is better than the DT, GNB or DNN algorithm with a significance level of 0.001, and better than the ERT algorithm with a significance level of 0.05. In addition, the superiority of the proposed MDA-GCNFTG has also been confirmed by experiments conducted on HMDD v3.0 (see

Table 4. The time and memory between the proposed novel GCN algorithm and the traditional GCN algorithm

	Algorithms	Balanced tasks			Unbalanced tasks		
		Tp	Tm	Td	Tp	Tm	Td
Time (s)	MDA-GCNFTG	13.45	8.6	4.9	13.05	10.95	14.55
	GCN	27.21	6.84	8.93	-	-	65.75
Memory (MB)	MDA-GCNFTG	2081	2109	2089	7275	7257	7109
	GCN	1651	1587	1509	-	-	7137

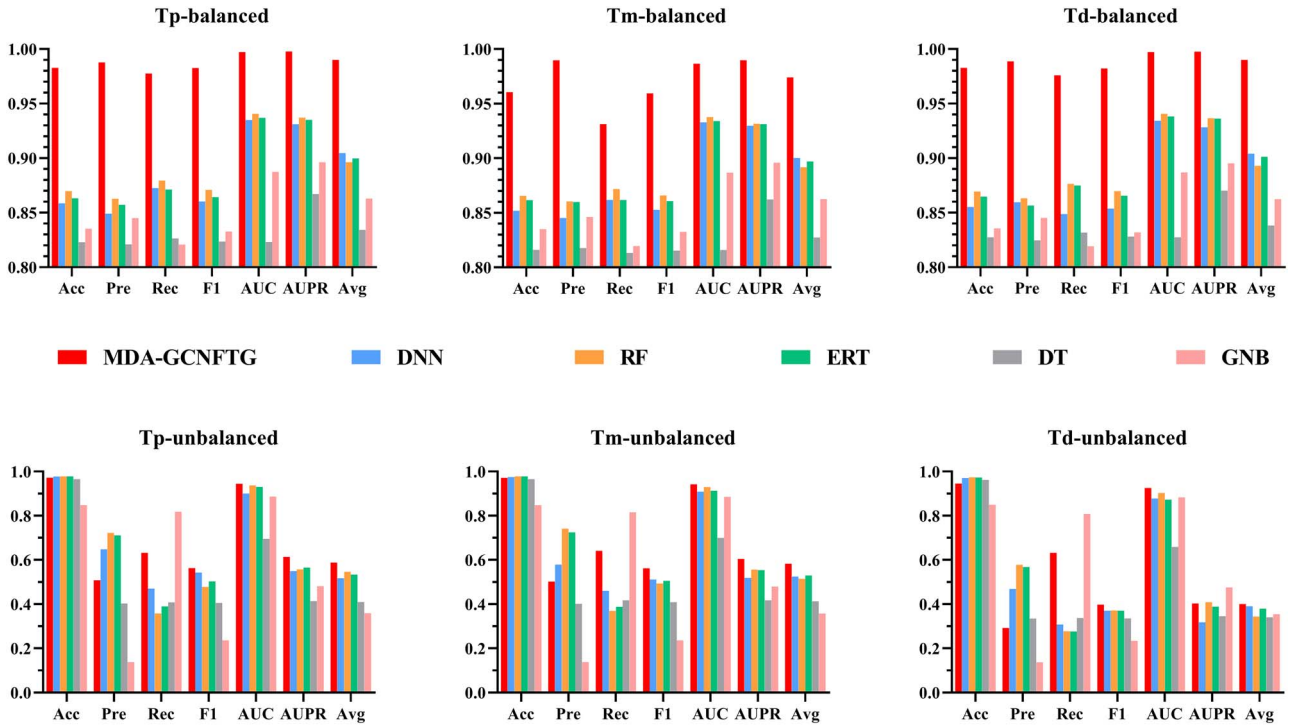

Figure 5. The comparison of the proposed MDA-GCNFTG method with some classic machine learning methods on six tasks, including DNN, RF, ERTs, DTs and GNBs. The Acc, Pre, Rec and F1 represent the accuracy, precision, recall and F1-score, respectively. For balanced tasks, the average value is calculated from accuracy, F1-score, AUC and AUPR. For unbalanced tasks, the average value is calculated from F1-score and AUPR.

Supplementary Tables S5–S7, see Supplementary Data available online at http://bib.oxfordjournals.org/).

Comparisons with the state-of-the-art methods

In order to further prove the superiority of the proposed MDA-GCNFTG method, we compare it with three state-of-the-art methods published after 2020, including GAMEDA [56], GBDT-LR [206] and DMA [241]. Note that all the following experiments are carried out under the same experimental conditions, including 5-fold CV, random seed and data partitioning strategy. We first verified the superiority of the MDA-GCNFTG method on the traditional task Tp (i.e. balanced Tp), and the results show that MDA-GCNFTG is better than the three state-of-the-art methods on almost all performance evaluation metrics (Figure 6).

Subsequently, we modified the code of GAMEDA, GBDT-LR and DMA to adapt to the other five tasks proposed in this study and compared them with the proposed MDA-GCNFTG method under the same experimental conditions. The results are shown in Figure 6. For balanced tasks, although MDA-GCNFTG has a slightly lower recall than GAMEDA and a slightly lower precision than DMA, overall, its performance is significantly better than these methods. In particular, recall and precision are usually a pair of contradictory performance metrics, and it is found that the precision of MDA-GCNFTG is much higher than that of GAMEDA, and the recall of MDA-GCNFTG is much higher than that of DMA. Therefore, the comprehensive performance metrics of recall and precision, which are F1-score and AUPR, must be considered and show the MDA-GCNFTG is higher than GAMEDA and DMA in these two performance metrics. Further, we explored the reason why GAMEDA has achieved such high recalls (that is, 1) and found that it predicted all samples as positive samples. And AUCs of the GAMEDA are 0.5 on these two tasks, which means that GAMEDA performs random classification, so other performance metrics seem to be meaningless. A similar phenomenon is appearing in the unbalanced task; thus, the F1-score and AUPR are considered to calculate the average value of performance evaluation metrics. And the proposed MDA-GCNFTG also achieved better performance for unbalanced tasks.

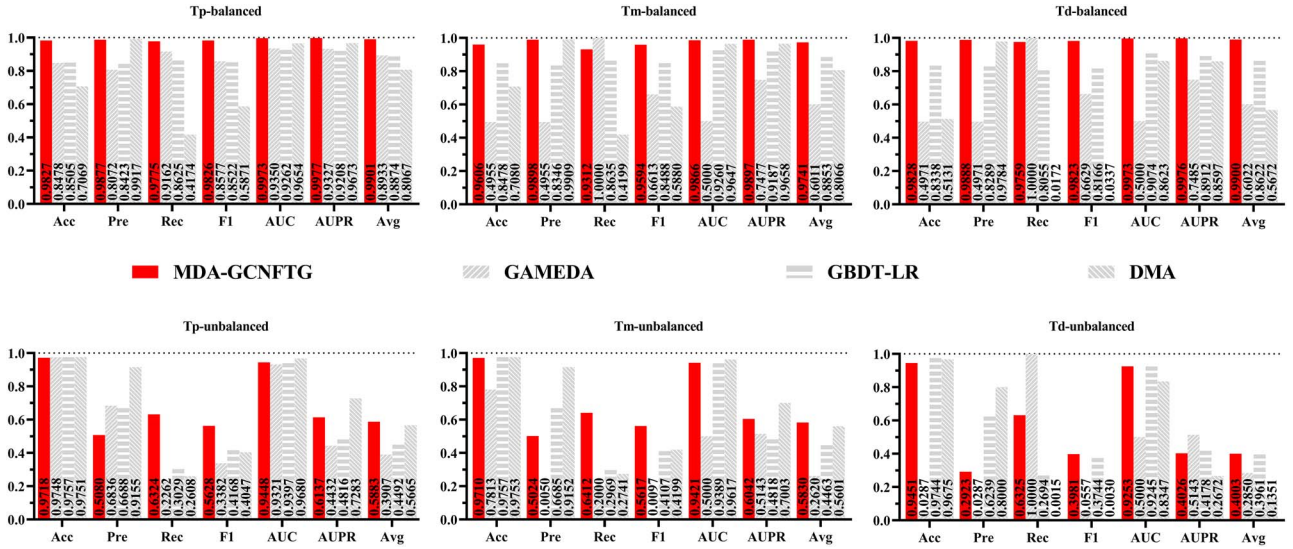

Figure 6. The comparison of the proposed MDA-GCNFTG with three state-of-the-art methods on six tasks under the same experimental conditions. Acc, Pre, Rec and F1 represent the accuracy, precision, recall and F1-score, respectively. For balanced tasks, the average value is calculated from accuracy, F1-score, AUC and AUPR. For unbalanced tasks, the average value is calculated from F1-score and AUPR.

Table 5. The summary of case studies for lung neoplasms, breast neoplasms, hsa-let-7a, hsa-let-7b and hsa-mir-1. Each case study is performed on HMDD v2.0 and an integrated data, which combine HMDD v3.2, miR2Disease and dbDEMC2 databases. The above two types of data are represented as 1 and 2 in the Data column, respectively. The Pos and Neg represent the number of positive and negative samples in the corresponding data. The FP, FN and TP are false positive, false negative and true positive of predicting results. The Top *n*/*m* column represents *n* of top *m* new MDAs are confirmed. The F1-score, AUC and AUPR are the performance of the corresponding data

	Data	Pos	Neg	FP	FN	TP	Top *n*/*m*	F1-score	AUC	AUPR
Lung	1	13	482	24	4	9		0.39	0.93	0.27
	2	28	467	9	4	24	27/28	0.79	0.97	0.77
Breast	1	24	471	17	9	15		0.54	0.94	0.38
	2	35	460	6	9	26	26/35	0.78	0.96	0.72
hsa-let-7a	1	45	338	12	12	33		0.73	0.93	0.81
	2	50	333	7	12	38	38/45	0.80	0.94	0.84
hsa-let-7b	1	38	345	25	8	30		0.65	0.94	0.83
	2	47	336	16	8	39	36/47	0.76	0.95	0.86
hsa-mir-1	1	46	337	25	14	32		0.62	0.94	0.69
	2	61	322	10	14	47	47/57	0.80	0.97	0.86

Figure 6 shows the significant superiority of MDA-GCNFTG over the state-of-the-art methods and also confirms the view that Tm and Td tasks are more difficult than Tp tasks and, at the same time, proves the robustness of the proposed MDA-GCNFTG method; that is, satisfactory results have been achieved on Tm, Td and unbalanced tasks. In addition, the superiority of the proposed MDA-GCNFTG has also been confirmed by experiments conducted on HMDD v3.0 (see Supplementary Tables S5–S7, see Supplementary Data available online at http://bib.oxfordjournals.org/).

Case studies

In order to further verify the performance of the proposed MDA-GCNFTG method on the MDA prediction problem, this study conducted two types of case studies for diseases and miRNAs, respectively. And this is the first time that case study for miRNA in the field of MDA prediction has been performed. On the other hand, we discuss the results of case studies in two types of data. The first data is the HMDD v2.0 database, which was used in this study. However, this database was proposed in 2014, and many new MDAs have been discovered during 7 years. Therefore, we integrated HMDD v3.2 [235], miR2Disease [242] and dbDEMC [243] database as the second data. It is worth noting that the second data is an update and expansion based on the first data.

For case studies on diseases, we chose lung neoplasms and breast neoplasms. Lung cancer is the most common fatal cancer with a high incidence. Although new drugs and treatments are being developed, the late presentation, poor prognosis and low cure rate still result in its high mortality rate. Many studies [244–246] have shown that some miRNAs can be used as biomarkers for lung cancer. Breast cancer is one of the most common cancers in women, and early detection and treatment can improve the prognosis of patients [247]. However, its complex clinical behavior and diverse histopathological patterns make huge challenge [247]. Evidence [247] shows that there is a close relationship between some miRNAs and breast cancer, so related miRNAs can be used as biomarkers to detect and prevent breast cancer.

Through extensive research on miRNAs, it has been determined that the let-7 miRNA family and hsa-mir-1 are related to a variety of human diseases [248–250]. Hsa-let-7a can induce

diseases with abnormal expression [251–253]. Hsa-let-7b is an important target of epigenetic mechanisms in various diseases [253–256]. Recent studies have also reported the association between hsa-mir-1 and various complex human diseases [257–259] and found the frequent methylation of hsa-mir-1 in colorectal cancer and believed that hsa-mir-1 played a tumor suppressor effect by controlling the expression of epithelial transition factor [260, 261].

The results and performances of the five case studies are listed in Table 5. It is clear that the case studies conducted on integrated data are convincing, and they all show satisfactory results, proving that the proposed MDA-GCNFTG method is capable of predicting the undiscovered potential MDA for new miRNAs and new diseases. The difference between the number of positive samples and TP on the two types of data also confirms this view and also reflects that the performance of the proposed MDA-GCNFTG method in this study is seriously underestimated.

Conclusion

MiRNAs have been shown to be closely related to numerous complex human diseases. Thus, predicting potential MDAs is essential to understand, prevent and treat diseases. This study designs a novel graph-construction strategy by using the *k*-NN algorithm and a novel GCN model based on graph sampling technology to do MDA prediction, that is, the MDA-GCNFTG method. Moreover, compared with other studies that only predict new MDAs based on balanced data, this study proposes two new experimental settings for predicting new miRNAs and predicting new diseases, and the above three experimental settings will be performed on balanced and unbalanced data, respectively. The results show that the proposed MDA-GCNFTG method has achieved satisfactory results on all six tasks and is superior to several classic machine learning algorithms and the most advanced MDA prediction methods. Moreover, we also conducted case studies for both miRNAs and diseases, which confirmed the effectiveness of our method. In the future, we hope to integrate other biological information and apply the data preprocessing technique on unbalanced data to obtain even better results. Moreover, most of the studies on MDA prediction (including this study) used the similarity-based measures derived from the known MDAs on the whole data set, which leads to the overoptimistic performance assessment of the current studies. In the next step, we will try to develop a more suitable feature representation method. For example, after dividing the training set and the test set, use the test sample masking method to first calculate the similarity among the training set samples and then use the *k*-NN algorithm to construct the similarity among test set samples.

Key Points

- This study designs a novel graph construction strategy by using the *k*-NN algorithm and a novel GCN model based on graph sampling technology to do MDA prediction, that is MDA-GCNFTG method.
- This study proposes two new experimental settings for predicting new miRNAs and predicting new diseases, and all experimental settings will be performed on balanced and unbalanced data, respectively.
- The results show that the proposed MDA-GCNFTG method has achieved satisfactory results on all six tasks and is superior to classic machine learning algorithms and the state-of-the-art MDA prediction methods.
- We also conducted case studies for both miRNAs and diseases, which confirmed the effectiveness of our method.

Supplementary data

Supplementary data are available online at https://academic.oup.com/bib.

Data availability

The data and source code are available from https://github.com/a96123155/MDA-GCNFTG.

Funding

National Science Foundation of China (32070662, 61832019, 32030063); Key Research Area Grant (2016YFA0501703) of the Ministry of Science and Technology of China, the Science and Technology Commission of Shanghai Municipality (19430750600); SJTU JiRLMDS Joint Research Fund and Joint Research Funds for Medical and Engineering and Scientific Research at Shanghai Jiao Tong University (ZH2018QNA41, YG2019GD01, YG2019ZDA12, YG2021ZD02). The computations were partially performed at the Pengcheng Lab. and the Center for High-Performance Computing, Shanghai Jiao Tong University.

References

1. Ambros V. microRNAs: tiny regulators with great potential. *Cell* 2001;**107**:823–6.
2. Ambros V. The function of animal MicroRNAs. *Nature* 2004;**431**:350–5.
3. Alshalalfa M, Alhajj R. Using context-specific effect of miRNAs to identify functional associations between miRNAs and gene signatures. *BMC Bioinformatics* 2013;**14**(Suppl 12):S1.
4. Bartel DP. MicroRNAs: target recognition and regulatory functions. *Cell* 2009;**136**:215–33.
5. Cheng AM, Byrom MW, Shelton J, *et al*. Antisense inhibition of human miRNAs and indications for an involvement of miRNA in cell growth and apoptosis. *Nucleic Acids Res* 2005;**33**:1290–7.
6. Karp X, Ambros V. Developmental biology. Encountering microRNAs in cell fate signaling. *Science* 2005;**310**:1288–9.
7. Miska EA. How microRNAs control cell division, differentiation and death. *Curr Opin Genet Dev* 2005;**15**:563–8.
8. Taganov KD, Boldin MP, Chang K, *et al*. NF-kappaB-dependent induction of microRNA miR-146, an inhibitor targeted to signaling proteins of innate immune responses. *Proc Natl Acad Sci U S A* 2006;**103**:12481–6.
9. Carleton M, Cleary MA, Linsley PS. MicroRNAS and cell cycle regulation. *Cell Cycle* 2007;**6**:2127–32.
10. Meng F, Henson R, Wehbe-Janek H, *et al*. MicroRNA-21 regulates expression of the PTEN tumor suppressor gene in human hepatocellular cancer. *Gastroenterology* 2007;**133**:647–58.

11. Taguchi Y. Inference of target gene regulation via miRNAs during cell senescence by using the MiRaGE server. *Aging Dis* 2012;**3**(4):301–6.
12. Griffiths-Jones S, Grocock RJ, van Dongen S, *et al*. miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Res* 2006;**34**:D140–4.
13. Meola N, Gennarino VA, Banfi S. microRNAs and genetic diseases. *PathoGenetics* 2009;**2**:7.
14. Urbich C, Kuehbacher A, Dimmeler S. Role of microRNAs in vascular diseases, inflammation, and angiogenesis. *Cardiovasc Res* 2008;**79**:581–8.
15. Latronico MVG, Catalucci D, Condorelli G. Emerging role of microRNAs in cardiovascular biology. *Circ Res* 2007;**101**:1225–36.
16. Small EM, Olson EN. Pervasive roles of microRNAs in cardiovascular biology. *Nature* 2011;**469**:336–42.
17. Hua S, Yun W, Zhiqiang Z, *et al*. A discussion of MicroRNAs in cancers. *Curr Bioinforma* 2014;**9**:453–62.
18. Miller TE, Ghoshal K, Ramaswamy B, *et al*. MicroRNA-221/222 confers tamoxifen resistance in breast cancer by targeting p27Kip1. *J Biol Chem* 2008;**283**:29897–903.
19. Madhavan D, Zucknick M, Wallwiener M, *et al*. Circulating miRNAs as surrogate markers for circulating tumor cells and prognostic markers in metastatic breast cancer. *Clin Cancer Res* 2012;**18**:5972–82.
20. Esquela-Kerscher A, Trang P, Wiggins JF, *et al*. The let-7 microRNA reduces tumor growth in mouse models of lung cancer. *Cell Cycle* 2008;**7**:759–64.
21. Zhu X, Li Y, Shen H, *et al*. miR-137 inhibits the proliferation of lung cancer cells by targeting Cdc42 and Cdk6. *FEBS Lett* 2013;**587**:73–81.
22. Chen RW, Bemis LT, Amato CM, *et al*. Truncation in CCND1 mRNA alters miR-16-1 regulation in mantle cell lymphoma. *Blood* 2008;**112**:822–9.
23. Lynam-Lennon N, Maher SG, Reynolds JV. The roles of microRNA in cancer and apoptosis. *Biol Rev Camb Philos Soc* 2009;**84**:55–71.
24. Chen X, Ba Y, Ma L, *et al*. Characterization of microRNAs in serum: a novel class of biomarkers for diagnosis of cancer and other diseases. *Cell Res* 2008;**18**:997–1006.
25. Zeng X, Zhang X, Zou Q. Integrative approaches for predicting microRNA function and prioritizing disease-related microRNA using biological interaction networks. *Brief Bioinform* 2016;**17**:193–203.
26. Calin GA, Croce CM. MicroRNA signatures in human cancers. *Nat Rev Cancer* 2006;**6**:857–66.
27. Chen X, Xie D, Zhao Q, *et al*. MicroRNAs and complex diseases: from experimental results to computational models. *Brief Bioinform* 2019;**20**:515–39.
28. Ravasz E, Somera AL, Mongru DA, *et al*. Hierarchical organization of modularity in metabolic networks. *Science* 2002;**297**:1551–5.
29. Barabási A, Gulbahce N, Loscalzo J. Network medicine: a network-based approach to human disease. *Nat Rev Genet* 2011;**12**:56–68.
30. Zhang S, Tong H, Xu J, *et al*. Graph convolutional networks: a comprehensive review. *Comput Soc Netw* 2019;**6**:1–23.
31. Jiang M, Li Z, Zhang S, *et al*. Drug-target affinity prediction using graph neural network and contact maps. *RSC Adv* 2020;**10**:20701–12.
32. Lim J, Ryu S, Park K, *et al*. Predicting drug-target interaction using a novel graph neural network with 3D structure-embedded graph representation. *J Chem Inf Model* 2019;**59**:3981–8.
33. Manoochehri HE, Pillai A, Nourani M Graph convolutional networks for predicting drug-protein interactions. In: *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. San Diego, CA, USA: IEEE, 345 E 47TH ST, NEW YORK, NY 10017 USA, 2019, pp. 1223–5.
34. Nguyen T, Le H, Quinn TP, *et al*. GraphDTA: predicting drug-target binding affinity with graph neural networks. *Bioinformatics* 2020;btaa921. doi: 10.1093/bioinformatics/btaa921.
35. Sun C, Xuan P, Zhang T, *et al*. Graph convolutional autoencoder and generative adversarial network-based method for predicting drug-target interactions. *IEEE/ACM Trans Comput Biol Bioinform* 2020;**PP**:1.
36. Zhao T, Hu Y, Valsdottir LR, *et al*. Identifying drug-target interactions based on graph convolutional network and deep neural network. *Brief Bioinform* 2021;**22**:2141–50.
37. Purkayastha S, Mondal I, Sarkar S, *et al*. Drug-drug interactions prediction based on drug embedding and graph autoencoder. In: *2019 IEEE 19th International Conference on Bioinformatics and Bioengineering (BIBE)*. Athens, GREECE: IEEE, 345 E 47TH ST, NEW YORK, NY 10017 USA, 2019, pp. 547–52.
38. Tran T, Kavuluru R, Kilicoglu H. Attention-gated graph convolutions for extracting drug interaction information from drug labels. *ACM Trans Comput Healthcare* 2021;**2**:10–1145.
39. Xiong WT, Li F, Yu H, *et al*. Extracting drug-drug interactions with a dependency-based graph convolution neural network. In: *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. San Diego, CA, USA: IEEE, 345 E 47TH ST, NEW YORK, NY 10017 USA, 2019, pp. 755–9.
40. Zhong Y, Chen X, Zhao Y, *et al*. Graph-augmented convolutional networks on drug-drug interactions prediction arXiv preprint arXiv:1912.03702 2019. 10 December 2019; preprint: not peer reviewed.
41. Aditya R, Saipradeep V, Thomas J, *et al*. Phenotype-driven gene prioritization for rare diseases using graph convolution on heterogeneous networks. *BMC Med Genet* 2018;**11**:57.
42. Han P, Yang P, Zhao P, *et al*. GCN-MF: disease-gene association identification by graph convolutional networks and matrix factorization. In: *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. Anchorage, AK, USA: ASSOC COMPUTING MACHINERY, 1515 BROADWAY, NEW YORK, NY 10036-9998 USA, 2019, pp. 705–13.
43. Singh V, Lio P. Towards probabilistic generative models harnessing graph neural networks for disease-gene prediction arXiv preprint arXiv:1907.05628 2019. 15 July 2019; preprint: not peer reviewed.
44. Zhang Y, Chen L, CIPHER-SC LS. Disease-gene association inference using graph convolution on a context-aware network with single-cell data. *IEEE/ACM Trans Comput Biol Bioinform* 2020;**PP**:1.
45. Kipf TN, Welling M. Semi-supervised classification with graph convolutional networks. In: *5th International Conference on Learning Representations (ICLR)*, Toulon, France, 2017.
46. Chen J, Zhu J, Song L. Stochastic training of graph convolutional networks with variance reduction. In: *35th International Conference on Machine Learning (ICML)*, Stockholm, Sweden, **2018**:941–949.
47. Chen J, Ma T, Xiao C. Fastgcn: fast learning with graph convolutional networks via importance sampling. In: *6th International Conference on Learning Representations (ICLR)*, Vancouver, BC, Canada, 2018.
48. Hamilton W, Ying Z, Leskovec J Inductive representation learning on large graphs. In: *Advances in Neural Information Processing Systems*. Long Beach, California, USA:

Neural Information Processing Systems (NIPS), 10010 NORTH TORREY PINES RD, LA JOLLA, CALIFORNIA 92037 USA, 2017, pp. 1024–34.
49. Huang W, Zhang T, Rong Y, *et al*. Adaptive sampling towards fast graph representation learning. In: *32nd Conference on Neural Information Processing Systems (NeurIPS)*, Montréal, Canada, **2018**:4563–4572.
50. Ying R, He R, Chen K, *et al*. Graph convolutional neural networks for web-scale recommender systems. In: *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. London, United Kingdom: ASSOC COMPUTING MACHINERY, 1515 BROADWAY, NEW YORK, NY 10036-9998 USA, 2018, pp. 974–83.
51. Zeng H, Zhou H, Srivastava A, *et al*. Graphsaint: graph sampling based inductive learning method. In: *8th International Conference on Learning Representations (ICLR)*, Addis Ababa, Ethiopia, 2020.
52. Wu Y, Cao N, Archambault D, *et al*. Evaluation of graph sampling: a visualization perspective. *IEEE Trans Vis Comput Graph* 2017;**23**:401–10.
53. Chiang W, Liu X, Si S, *et al*. Cluster-GCN: an efficient algorithm for training deep and large graph convolutional networks. In: *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. Anchorage, AK, USA: ASSOC COMPUTING MACHINERY, 1515 BROADWAY, NEW YORK, NY 10036-9998 USA, 2019, pp. 257–66.
54. Zeng H, Zhou H, Srivastava A, *et al*. Accurate, efficient and scalable graph embedding. In: *2019 IEEE International Parallel and Distributed Processing Symposium (IPDPS)*, IEEE, 2019, pp. 462–71.
55. Li J, Chen X, Huang Q, *et al*. Seq-SymRF: a random forest model predicts potential miRNA-disease associations based on information of sequences and clinical symptoms. *Sci Rep* 2020;**10**:17901.
56. Li Z, Li J, Nie R, *et al*. A graph auto-encoder model for miRNA-disease associations prediction. *Brief Bioinform* 2020;bbaa240. doi: 10.1093/bib/bbaa240.
57. Ding Y, Tian L, Lei X, *et al*. Variational graph auto-encoders for miRNA-disease association prediction. *Methods* 2020;**S1046–2023**:30164.
58. Li J, Li Z, Nie R, *et al*. FCGCNMDA: predicting miRNA-disease associations by applying fully connected graph convolutional networks. *Mol Gen Genomics* 2020;**295**:1197–209.
59. Rout N, Mishra D, Mallick MK. Handling imbalanced data: a survey, international proceedings on advances in soft computing. *Int Syst Appl* 2018;**628**:431–43.
60. Altman NS. An introduction to kernel and nearest-neighbor nonparametric regression. *Am Stat* 1992;**46**:175–85.
61. Fix E. *Discriminatory analysis: nonparametric discrimination: consistency properties, Report No.4*. Randolph Field, TX: USAF School of Aviation Medicine, 1951.
62. Chen X, Yan CC, Zhang X, *et al*. WBSMDA: within and between score for MiRNA-disease association prediction. *Sci Rep* 2016;**6**:21106.
63. Huang Z, Liu L, Gao Y, *et al*. Benchmark of computational methods for predicting microRNA-disease associations. *Genome Biol* 2019;**20**:202.
64. Jiang Q, Hao Y, Wang G, *et al*. Prioritization of disease microRNAs through a human phenome-microRNAome network. *BMC Syst Biol* 2010;**4**(Suppl 1):S2.
65. Jiang Y, Liu B, Yu L, *et al*. Predict MiRNA-disease association with collaborative filtering. *Neuroinformatics* 2018;**16**:363–72.
66. Chen X, Niu YW, Wang GH, *et al*. MKRMDA: multiple kernel learning-based Kronecker regularized least squares for MiRNA-disease association prediction. *J Transl Med* 2017;**15**:251.
67. Lan W, Wang J, Li M, *et al*. Predicting microRNA-disease associations by integrating multiple biological information. In: *IEEE International Conference on Bioinformatics and Biomedicine*. Washington, DC, USA: IEEE, 345 E 47TH ST, NEW YORK, NY 10017 USA, 2015, pp. 183–8.
68. Shi H, Xu J, Zhang G, *et al*. Walking the interactome to identify human miRNA-disease associations through the functional link between miRNA targets and disease genes. *BMC Syst Biol* 2013;**7**:101.
69. Mørk S, Pletscher-Frankild S, Caro AP, *et al*. Protein-driven inference of miRNA-disease associations. *Bioinformatics* 2014;**30**:392–7.
70. Xu C, Ping Y, Li X, *et al*. Prioritizing candidate disease miRNAs by integrating phenotype associations of multiple diseases with matched miRNA and mRNA expression profiles. *Mol BioSyst* 2014;**10**:2800–9.
71. Xu P, Wu Q, Lu D, *et al*. A systematic study of critical miRNAs on cells proliferation and apoptosis by the shortest path. *BMC Bioinformatics* 2020;**21**:396.
72. Wang C, Sun K, Wang J, *et al*. Data fusion-based algorithm for predicting miRNA-disease associations. *Comput Biol Chem* 2020;**88**:107357.
73. Liu M, Yang J, Wang J, *et al*. Predicting miRNA-disease associations using a hybrid feature representation in the heterogeneous network. *BMC Med Genet* 2020;**13**:153.
74. Le D, TTH T. RWRMTN: a tool for predicting disease-associated microRNAs based on a microRNA-target gene network. *BMC Bioinformatics* 2020;**21**:244.
75. Yu L, Shen X, Zhong D, *et al*. Three-layer heterogeneous network combined with unbalanced random walk for miRNA-disease association prediction. *Front Genet* 2019;**10**:1316.
76. Yu DL, Ma YL, Yu ZG. Inferring microRNA-disease association by hybrid recommendation algorithm and unbalanced bi-random walk on heterogeneous network. *Sci Rep* 2019;**9**:2474.
77. Ma Y, He T, Ge L, *et al*. MiRNA-disease interaction prediction based on kernel neighborhood similarity and multi-network bidirectional propagation. *BMC Med Genet* 2019;**12**:185.
78. Liu W, Cui Z, Zan X. Identifying cancer-related microRNAs based on subpathways. *IET Syst Biol* 2018;**12**:273–8.
79. Li X, Lin Y, Gu C, *et al*. SRMDAP: SimRank and density-based clustering recommender model for miRNA-disease association prediction. *Biomed Res Int* 2018;**2018**:5747489.
80. Ding P, Luo J, Liang C, *et al*. Human disease MiRNA inference by combining target information based on heterogeneous manifolds. *J Biomed Inform* 2018;**80**:26–36.
81. He BS, Qu J, Chen M. Prediction of potential disease-associated microRNAs by composite network based inference. *Sci Rep* 2018;**8**:15813.
82. Chen X, Zhang DH, You ZH. A heterogeneous label propagation approach to explore the potential associations between miRNA and disease. *J Transl Med* 2018;**16**:348.
83. Chen X, Qu J, Yin J. TLHNMDA: triple layer heterogeneous network based inference for MiRNA-disease association prediction. *Front Genet* 2018;**9**:234.
84. Peng W, Lan W, Zhong J, *et al*. A novel method of predicting microRNA-disease associations based on microRNA, disease, gene and environment factor networks. *Methods* 2017;**124**:69–77.

85. Peng W, Lan W, Yu Z, *et al*. A framework for integrating multiple biological networks to predict MicroRNA-disease associations. *IEEE Trans Nanobioscience* 2017;**16**:100–7.
86. Peng H, Lan C, Zheng Y, *et al*. Cross disease analysis of co-functional microRNA pairs on a reconstructed network of disease-gene-microRNA tripartite. *BMC Bioinformatics* 2017;**18**:193.
87. Pallez D, Gardes J, Pasquier C. Prediction of miRNA-disease associations using an evolutionary tuned latent semantic analysis. *Sci Rep* 2017;**7**:10548.
88. Liu Y, Zeng X, He Z, *et al*. Inferring microRNA-disease associations by random walk on a heterogeneous network with multiple data sources. *IEEE/ACM Trans Comput Biol Bioinform* 2017;**14**:905–15.
89. Le D, Verbeke L, Son LH, *et al*. Random walks on mutual microRNA-target gene interaction network improve the prediction of disease-associated microRNAs. *BMC Bioinformatics* 2017;**18**:479.
90. Huang YA, You ZH, Li LP, *et al*. EPMDA: an expression-profile based computational model for microRNA-disease association prediction. *Oncotarget* 2017;**8**:87033–43.
91. Shi H, Zhang G, Zhou M, *et al*. Integration of multiple genomic and phenotype data to infer novel miRNA-disease associations. *PLoS One* 2016;**11**:e148521.
92. Qin GM, Li RY, Zhao XM. Identifying disease associated miRNAs based on protein domains. *IEEE/ACM Trans Comput Biol Bioinform* 2016;**13**:1027–35.
93. Zhao XM, Liu KQ, Zhu G, *et al*. Identifying cancer-related microRNAs based on gene expression data. *Bioinformatics* 2015;**31**:1226–34.
94. Yuan D, Cui X, Wang Y, *et al*. Enrichment analysis identifies functional MicroRNA-disease associations in humans. *PLoS One* 2015;**10**:e136285.
95. Ha J, Kim H, Yoon Y, *et al*. A method of extracting disease-related microRNAs through the propagation algorithm using the environmental factor based global miRNA network. *Biomed Mater Eng* 2015;**26**(Suppl 1):S1763–72.
96. Xiao Q, Dai J, Luo J, *et al*. Multi-view manifold regularized learning-based method for prioritizing candidate disease miRNAs. *Knowl-Based Syst* 2019;**175**:118–29.
97. Lv H, Li J, Zhang S, *et al*. Meta-path based MiRNA-disease association prediction. In: *Database Systems for Advanced Applications*. Cham: Springer International Publishing, 2019, 34–48.
98. Ding L, Wang M, Sun D, *et al*. A novel method for identifying potential disease-related miRNAs via a disease–miRNA–target heterogeneous network. *Mol BioSyst* 2017;**13**:2328–2337.
99. Peng W, Lan W, Yu Z, *et al*. Predicting MicroRNA-disease associations by random walking on multiple networks. In: *International Symposium on Bioinformatics Research and Applications*. Minsk, BYELARUS: SPRINGER INTERNATIONAL PUBLISHING AG, GEWERBESTRASSE 11, CHAM, CH-6330, SWITZERLAND, 2016, 127–35.
100. Chen H, Zhang Z. Similarity-based methods for potential human microRNA-disease association prediction. *BMC Med Genet* 2013;**6**:12.
101. Zhang L, Liu B, Li Z, *et al*. Predicting MiRNA-disease associations by multiple meta-paths fusion graph embedding model. *BMC Bioinformatics* 2020;**21**:470.
102. Huang F, Yue X, Xiong Z, *et al*. Tensor decomposition with relational constraints for predicting multiple types of microRNA-disease associations. *Brief Bioinform* 2020;bbaa140. doi: 10.1093/bib/bbaa140.
103. Yan F, Zheng Y, Jia W, *et al*. MAMDA: inferring microRNA-disease associations with manifold alignment. *Comput Biol Med* 2019;**110**:156–63.
104. Qu J, Chen X, Yin J, *et al*. Prediction of potential miRNA-disease associations using matrix decomposition and label propagation. *Knowl-Based Syst* 2019;**186**:104963.
105. Mao G, Wang SL, Zhang W. Prediction of potential associations between MicroRNA and disease based on Bayesian probabilistic matrix factorization model. *J Comput Biol* 2019;**26**:1030–9.
106. Ding T, Gao J, Zhu S, *et al*. Predicting microRNA-disease association based on microRNA structural and functional similarity network. *Quant Biol* 2019;**7**:138–46.
107. Chen Q, Zhao Z, Lan W, *et al*. Predicting miRNA-disease interaction based on recommend method. *Inf Discov Deliv* 2019;**48**:35–40.
108. Shao B, Liu B, Yan C. SACMDA: MiRNA-disease association prediction with short acyclic connections in heterogeneous graph. *Neuroinformatics* 2018;**16**:373–82.
109. Luo J, Ding P, Liang C, *et al*. Semi-supervised prediction of human miRNA-disease association based on graph regularization framework in heterogeneous networks. *Neurocomputing* 2018;**294**:29–38.
110. Chen L, Liu B, Yan C. DPFMDA: distributed and privatized framework for miRNA-disease association prediction. *Pattern Recogn Lett* 2018;**109**:4–11.
111. Jiang ZC, Shen Z, Bao W A novel computational method for MiRNA-disease association prediction. In: *International Conference on Intelligent Computing*. Liverpool, ENGLAND: SPRINGER INTERNATIONAL PUBLISHING AG, GEWERBESTRASSE 11, CHAM, CH-6330, SWITZERLAND, 2017, pp. 539–47.
112. Zou Q, Li J, Song L, *et al*. Similarity computation strategies in the microRNA-disease network: a survey. *Brief Funct Genomics* 2016;**15**:55–64.
113. Zhao Y, Chen X, Yin J. A novel computational method for the identification of potential miRNA-disease association based on symmetric non-negative matrix factorization and Kronecker regularized least square. *Front Genet* 2018;**9**:324.
114. Zhao Q, Xie D, Liu H, *et al*. SSCMDA: spy and super cluster strategy for MiRNA-disease association prediction. *Oncotarget* 2018;**9**:1826–42.
115. Zhao H, Kuang L, Feng X, *et al*. A novel approach based on a weighted interactive network to predict associations of MiRNAs and diseases. *Int J Mol Sci* 2018;**20**:110.
116. Yang Y, Fu X, Qu W, *et al*. MiRGOFS: a GO-based functional similarity measurement for miRNAs, with applications to the prediction of miRNA subcellular localization and miRNA-disease association. *Bioinformatics* 2018;**34**:3547–56.
117. Xiao Q, Luo J, Liang C, *et al*. A graph regularized non-negative matrix factorization method for identifying microRNA-disease associations. *Bioinformatics* 2018;**34**:239–48.
118. Sun Y, Zhu Z, You ZH, *et al*. FMSM: a novel computational model for predicting potential miRNA biomarkers for various human diseases. *BMC Syst Biol* 2018;**12**:121.
119. Qu Y, Zhang H, Lyu C, *et al*. LLCMDA: a novel method for predicting miRNA gene and disease relationship based on locality-constrained linear coding. *Front Genet* 2018;**9**:576.
120. Qu Y, Zhang H, Liang C, *et al*. SNMDA: a novel method for predicting microRNA-disease associations based on sparse neighbourhood. *J Cell Mol Med* 2018;**22**:5109–20.

121. Peng LH, Sun CN, Guan NN, *et al*. HNMDA: heterogeneous network-based miRNA-disease association prediction. *Mol Gen Genomics* 2018;**293**:983–95.
122. Liang C, Yu S, Wong KC, *et al*. A novel semi-supervised model for miRNA-disease association prediction based on ℓ1-norm graph. *J Transl Med* 2018;**16**:357.
123. Li G, Luo J, Xiao Q, *et al*. Predicting microRNA-disease associations using label propagation based on linear neighborhood similarity. *J Biomed Inform* 2018;**82**:169–77.
124. Jiang L, Xiao Y, Ding Y, *et al*. FKL-spa-LapRLS: an accurate method for identifying human microRNA-disease association. *BMC Genomics* 2018;**19**:911.
125. Jiang L, Ding Y, Tang J, *et al*. MDA-SKF: similarity kernel fusion for accurately discovering miRNA-disease association. *Front Genet* 2018;**9**:618.
126. He BS, Qu J, Zhao Q. Identifying and exploiting potential miRNA-disease associations with Neighborhood regularized logistic matrix factorization. *Front Genet* 2018;**9**:303.
127. Chen X, Yang JR, Guan NN, *et al*. GRMDA: graph regression for MiRNA-disease association prediction. *Front Physiol* 2018;**9**:92.
128. Chen X, Wang LY, Huang L. NDAMDA: network distance analysis for MiRNA-disease association prediction. *J Cell Mol Med* 2018;**22**:2884–95.
129. Chen X, Wang L, Qu J, *et al*. Predicting miRNA-disease association based on inductive matrix completion. *Bioinformatics* 2018;**34**:4256–65.
130. Chen X, Guan NN, Li JQ, *et al*. GIMDA: graphlet interaction-based MiRNA-disease association prediction. *J Cell Mol Med* 2018;**22**:1548–61.
131. Chen X, Cheng JY, Yin J. Predicting microRNA-disease associations using bipartite local models and hubness-aware regression. *RNA Biol* 2018;**15**:1192–205.
132. Chen M, Liao B, Li Z. Global similarity method based on a two-tier random walk for the prediction of microRNA-disease association. *Sci Rep* 2018;**8**:6481.
133. Yu H, Chen X, Lu L. Large-scale prediction of microRNA-disease associations by combinatorial prioritization algorithm. *Sci Rep* 2017;**7**:43792.
134. Peng L, Peng M, Liao B, *et al*. Improved low-rank matrix recovery method for predicting miRNA-disease association. *Sci Rep* 2017;**7**:6007.
135. Peng L, Chen Y, Ma N, *et al*. NARRMDA: negative-aware and rating-based recommendation algorithm for miRNA-disease association prediction. *Mol BioSyst* 2017;**13**:2650–9.
136. Luo J, Xiao Q. A novel approach for predicting microRNA-disease associations by unbalanced bi-random walk on heterogeneous network. *J Biomed Inform* 2017;**66**:194–203.
137. Luo J, Ding P, Liang C, *et al*. Collective prediction of disease-associated miRNAs based on transduction learning. *IEEE/ACM Trans Comput Biol Bioinform* 2017;**14**:1468–75.
138. Li X, Lin Y, Gu C. A network similarity integration method for predicting microRNA-disease associations. *RSC Adv* 2017;**7**:32216–24.
139. Chen X, Niu YW, Wang GH, *et al*. HAMDA: hybrid approach for MiRNA-disease association prediction. *J Biomed Inform* 2017;**76**:50–8.
140. Gu C, Liao B, Li X, *et al*. Network consistency projection for human miRNA-disease associations inference. *Sci Rep* 2016;**6**:36054.
141. Chen M, Lu X, Liao B, *et al*. Uncover miRNA-disease association by exploiting global network similarity. *PLoS One* 2016;**11**:e166509.
142. Liao B, Ding S, Chen H, *et al*. Identifying human microRNA-disease associations by a new diffusion-based method. *J Bioinforma Comput Biol* 2015;**13**:1550014.
143. Le D. Network-based ranking methods for prediction of novel disease associated microRNAs. *Comput Biol Chem* 2015;**58**:139–48.
144. Han K, Xuan P, Ding J, *et al*. Prediction of disease-related microRNAs by incorporating functional similarity and common association information. *Genet Mol Res* 2014;**13**:2009–19.
145. Chen H, Zhang Z. Prediction of associations between OMIM diseases and microRNAs by random walk on OMIM disease similarity network. *Sci World J* 2013;**2013**:204658.
146. Chen X, Liu MX, Yan GY. RWRMDA: predicting novel human microRNA-disease associations. *Mol BioSyst* 2012;**8**:2792–8.
147. Xuan P, Han K, Guo Y, *et al*. Prediction of potential disease-associated microRNAs based on random walk. *Bioinformatics* 2015;**31**:1805–15.
148. Sun D, Li A, Feng H, *et al*. NTSMDA: prediction of miRNA-disease associations by integrating network topological similarity. *Mol BioSyst* 2016;**12**:2224–32.
149. Chen X, Yan CC, Zhang X, *et al*. HGIMDA: heterogeneous graph inference for miRNA-disease association prediction. *Oncotarget* 2016;**7**:65257–69.
150. Ding P, Luo J, Xiao Q, *et al*. A path-based measurement for human miRNA functional similarities using miRNA-disease associations. *Sci Rep* 2016;**6**:32533.
151. You Z, Huang Z, Zhu Z, *et al*. PBMDA: a novel and effective path-based computational model for miRNA-disease association prediction. *PLoS Comput Biol* 2017;**13**:e1005455.
152. Zheng K, You ZH, Wang L, *et al*. DBMDA: a unified embedding for sequence-based miRNA similarity measure with applications to predict and validate miRNA-disease association. *Mol Ther Nucleic Acids* 2020;**19**:602–11.
153. Gao Z, Wang Y, Wu Q, *et al*. Graph regularized $L_{2,1}$-nonnegative matrix factorization for miRNA-disease association prediction. *BMC Bioinformatics* 2020;**21**:61.
154. Chen X, Li S, Yin J, *et al*. Potential miRNA-disease association prediction based on kernelized Bayesian matrix factorization. *Genomics* 2020;**112**:809–19.
155. Zhu X, Wang X, Zhao H, *et al*. BHCMDA: a new biased heat conduction based method for potential MiRNA-disease association prediction. *Front Genet* 2020;**11**:384.
156. Chen X, Sun L, Zhao Y. NCMCMDA: miRNA-disease association prediction through neighborhood constraint matrix completion. *Brief Bioinform* 2021;**22**:485–96.
157. Zhang Y, Chen M, Cheng X, *et al*. MSFSP: a novel miRNA–disease association prediction model by federating multiple-similarities fusion and space projection. *Front Genet* 2020;**11**:389.
158. Zeng X, Wang W, Deng G, *et al*. Prediction of potential disease-associated MicroRNAs by using neural networks. *Mol Ther Nucleic Acids* 2019;**16**:566–75.
159. Zeng X, Liu L, Lü L, *et al*. Prediction of potential disease-associated microRNAs using structural perturbation method. *Bioinformatics* 2018;**34**:2425–32.
160. Zhang X, Zou Q, Rodriguez-Paton A, *et al*. Meta-path methods for prioritizing candidate disease miRNAs. *IEEE/ACM Trans Comput Biol Bioinform* 2019;**16**:283–91.
161. Chen X, Jiang Z, Xie D, *et al*. A novel computational model based on super-disease and miRNA for potential miRNA-disease association prediction. *Mol BioSyst* 2017;**13**:1202–12.

162. Chen X, Xie D, Wang L, *et al*. BNPMDA: bipartite network projection for MiRNA-disease association prediction. *Bioinformatics* 2018;**34**:3178–86.
163. Chen X, Yin J, Qu J, *et al*. MDHGI: matrix decomposition and heterogeneous graph inference for miRNA-disease association prediction. *PLoS Comput Biol* 2018;**14**:e1006418.
164. Gong Y, Niu Y, Zhang W, *et al*. A network embedding-based multiple information integration method for the MiRNA-disease association prediction. *BMC Bioinformatics* 2019;**20**:468.
165. Shen Z, Zhang Y, Han K, *et al*. miRNA-disease association prediction with collaborative matrix factorization. *Complexity* 2017;**2017**:1–9.
166. You Z, Wang L, Chen X, *et al*. PRMDA: personalized recommendation-based MiRNA-disease association prediction. *Oncotarget* 2017;**8**:85568–83.
167. Cao B, Deng S, Qin H, *et al*. Inferring MicroRNA-disease associations based on the identification of a functional module. *J Comput Biol* 2021;**28**:33–42.
168. Wu TR, Yin MM, Jiao CN, *et al*. MCCMF: collaborative matrix factorization based on matrix completion for predicting miRNA-disease associations. *BMC Bioinformatics* 2020;**21**:454.
169. Wu Q, Wang Y, Gao Z, *et al*. MSCHLMDA: multi-similarity based combinative hypergraph learning for predicting MiRNA-disease association. *Front Genet* 2020;**11**:354.
170. Wang L, Chen Y, Zhang N, *et al*. QIMCMDA: MiRNA-disease association prediction by q-kernel information and matrix completion. *Front Genet* 2020;**11**:594796.
171. Sun P, Yang S, Cao Y, *et al*. Prediction of potential associations between miRNAs and diseases based on matrix decomposition. *Front Genet* 2020;**11**:598185.
172. Li J, Liu Y, Zhang Z, *et al*. PmDNE: prediction of miRNA-disease association based on network embedding and network similarity analysis. *Biomed Res Int* 2020;**2020**: 6248686.
173. Li H, Guo Y, Cai M, *et al*. MicroRNA-disease association prediction by matrix tri-factorization. *BMC Genomics* 2020;**21**:617.
174. Ha J, Park C, Park C, *et al*. Improved prediction of miRNA-disease associations based on matrix completion with network regularization. *Cell* 2020;**9**:881.
175. Guan NN, Wang CC, Zhang L, *et al*. In silico prediction of potential miRNA-disease association using an integrative bioinformatics approach based on kernel fusion. *J Cell Mol Med* 2020;**24**:573–87.
176. Chen H, Guo R, Li G, *et al*. Comparative analysis of similarity measurements in miRNAs with applications to miRNA-disease association predictions. *BMC Bioinformatics* 2020;**21**:176.
177. Zhang W, Li Z, Guo W, *et al*. A fast linear neighborhood similarity-based network link inference method to predict microRNA-disease associations. *IEEE/ACM Trans Comput Biol Bioinform* 2021;**18**:405–15.
178. Yu SP, Liang C, Xiao Q, *et al*. MCLPMDA: a novel method for miRNA-disease association prediction based on matrix completion and label propagation. *J Cell Mol Med* 2019;**23**:1427–38.
179. Yan C, Wang J, Ni P, *et al*. DNRLMF-MDA: predicting microRNA-disease associations based on similarities of microRNAs and diseases. *IEEE/ACM Trans Comput Biol Bioinform* 2019;**16**:233–43.
180. Xuan P, Zhang Y, Zhang T, *et al*. Predicting miRNA-disease associations by incorporating projections in low-dimensional space and local topological information. *Genes (Basel)* 2019;**10**:685.
181. Xuan P, Li L, Zhang T, *et al*. Prediction of disease-related microRNAs through integrating attributes of microRNA nodes and multiple kinds of connecting edges. *Molecules* 2019;**24**:3099.
182. Xie G, Fan Z, Sun Y, *et al*. WBNPMD: weighted bipartite network projection for microRNA-disease association prediction. *J Transl Med* 2019;**17**:322.
183. Pan Z, Zhang H, Liang C, *et al*. Self-weighted multi-kernel multi-label learning for potential miRNA-disease association prediction. *Mol Ther Nucleic Acids* 2019;**17**:414–23.
184. Liang C, Yu S, Luo J. Adaptive multi-view multi-label learning for identifying disease-associated candidate miRNAs. *PLoS Comput Biol* 2019;**15**:e1006931.
185. Li X, Lin Y, Gu C, *et al*. FCMDAP: using miRNA family and cluster information to improve the prediction accuracy of disease related miRNAs. *BMC Syst Biol* 2019;**13**:26.
186. Gao YL, Cui Z, Liu JX, *et al*. NPCMF: nearest profile-based collaborative matrix factorization method for predicting miRNA-disease associations. *BMC Bioinformatics* 2019;**20**:353.
187. Gao MM, Cui Z, Gao YL, *et al*. Dual-network sparse graph regularized matrix factorization for predicting miRNA-disease associations. *Mol Omics* 2019;**15**:130–7.
188. Cui Z, Liu JX, Gao YL, *et al*. RCMF: a robust collaborative matrix factorization method to predict miRNA-disease associations. *BMC Bioinformatics* 2019;**20**:686.
189. Chen M, Zhang Y, Li A, *et al*. Bipartite heterogeneous network method based on co-neighbor for MiRNA-disease association prediction. *Front Genet* 2019;**10**:385.
190. Che K, Guo M, Wang C, *et al*. Predicting MiRNA-disease association by latent feature extraction with positive samples. *Genes (Basel)* 2019;**10**:80.
191. Zhong Y, Xuan P, Wang X, *et al*. A non-negative matrix factorization based method for predicting disease-associated miRNAs in miRNA-disease bilayer network. *Bioinformatics* 2018;**34**:267–77.
192. Chen X, Clarence Yan C, Zhang X, *et al*. RBMMMDA: predicting multiple types of disease-microRNA associations. *Sci Rep* 2015;**5**:13877.
193. Zhang L, Ai H, Chen W, *et al*. CarcinoPred-EL: novel models for predicting the carcinogenicity of chemicals using molecular fingerprints and ensemble learning methods. *Sci Rep* 2017;**7**:2118.
194. Xu J, Li C, Lv J, *et al*. Prioritizing candidate disease miRNAs by topological features in the miRNA target-dysregulated network: case study of prostate cancer. *Mol Cancer Ther* 2011;**10**:1857–66.
195. Xuan P, Han K, Guo M, *et al*. Prediction of microRNAs associated with human diseases based on weighted k most similar neighbors. *PLoS One* 2013;**8**:e70204.
196. Chen X, Yan GY. Semi-supervised learning for potential human microRNA-disease associations inference. *Sci Rep* 2014;**4**:5501.
197. Pasquier C, Gardès J. Prediction of miRNA-disease associations with a vector space model. *Sci Rep* 2016;**6**:27036.
198. Jian-Qiang L, Zhi-Hao R, Xing C, *et al*. MCMDA: matrix completion for MiRNA-disease association prediction. *Oncotarget* 2017;**8**:21187.

199. Chen X, Wu QF, Yan GY. RKNNMDA: ranking-based KNN for MiRNA-disease association prediction. *RNA Biol* 2017;**14**:952–62.
200. Luo J, Xiao Q, Liang C, *et al*. Predicting MicroRNA-disease associations using Kronecker regularized least squares based on heterogeneous omics data. *IEEE Access* 2017;**5**:2503–13.
201. Liu Y, Wang SL, Zhang JF, *et al*. A neural collaborative filtering method for identifying miRNA-disease associations. *Neurocomputing* 2021;**422**:176–85.
202. Chen X, Gong Y, Zhang D, *et al*. DRMDA: deep representations-based miRNA-disease association prediction. *J Cell Mol Med* 2018;**22**:472–85.
203. Chen X, Huang L. LRSSLMDA: Laplacian regularized sparse subspace learning for MiRNA-disease association prediction. *PLoS Comput Biol* 2017;**13**:e1005912.
204. Chen X, Huang L, Xie D, *et al*. EGBMMDA: extreme gradient boosting machine for MiRNA-disease association prediction. *Cell Death Dis* 2018;**9**:3.
205. Chen X, Zhou Z, Zhao Y. ELLPMDA: ensemble learning and link prediction for miRNA-disease association prediction. *RNA Biol* 2018;**15**:807–18.
206. Zhou S, Wang S, Wu Q, *et al*. Predicting potential miRNA-disease associations by combining gradient boosting decision tree with logistic regression. *Comput Biol Chem* 2020;**85**:107200.
207. Peng LH, Zhou LQ, Chen X, *et al*. A computational study of potential miRNA-disease association inference based on ensemble learning and kernel ridge regression. *Front Bioeng Biotechnol* 2020;**8**:40.
208. Ha J, Park C, Park C, *et al*. IMIPMF: inferring miRNA-disease interactions using probabilistic matrix factorization. *J Biomed Inform* 2020;**102**:103358.
209. Ding Y, Wang F, Lei X, *et al*. Deep belief network-based matrix factorization model for MicroRNA-disease associations prediction. *Evol Bioinformatics Online* 2020;**16**:1612666011.
210. Ding Y, Jiang L, Tang J, *et al*. Identification of human microRNA-disease association via hypergraph embedded bipartite local model. *Comput Biol Chem* 2020;**89**:107369.
211. Chen X, Li TH, Zhao Y, *et al*. Deep-belief network for predicting potential miRNA-disease associations. *Brief Bioinform* 2020;bbaa186. doi: 10.1093/bib/bbaa186.
212. Zhao Y, Chen X, Yin J. Adaptive boosting-based computational model for predicting potential miRNA-disease associations. *Bioinformatics* 2019;**35**:4730–8.
213. Zhang L, Chen X, Yin J. Prediction of potential miRNA-disease associations through a novel unsupervised deep learning framework with Variational autoencoder. *Cell* 2019;**8**:1040.
214. Yao D, Zhan X, Kwoh CK. An improved random forest-based computational model for predicting novel miRNA-disease associations. *BMC Bioinformatics* 2019;**20**:624.
215. Xuan P, Sun H, Wang X, *et al*. Inferring the disease-associated miRNAs based on network representation learning and convolutional neural networks. *Int J Mol Sci* 2019;**20**:3648.
216. Wu M, Yang Y, Wang H, *et al*. IMPMD: an integrated method for predicting potential associations between miRNAs and diseases. *Curr Genomics* 2019;**20**:581–91.
217. Wang L, You ZH, Chen X, *et al*. LMTRDA: using logistic model tree to predict MiRNA-disease associations by fusing multi-source information of sequences and similarities. *PLoS Comput Biol* 2019;**15**:e1006865.
218. Wang CC, Chen X, Yin J, *et al*. An integrated framework for the identification of potential miRNA-disease association based on novel negative samples extraction strategy. *RNA Biol* 2019;**16**:257–69.
219. Peng J, Hui W, Li Q, *et al*. A learning-based framework for miRNA-disease association identification using neural networks. *Bioinformatics* 2019;**35**:4364–71.
220. Niu YW, Wang GH, Yan GY, *et al*. Integrating random walk and binary regression to identify novel miRNA-disease association. *BMC Bioinformatics* 2019;**20**:59.
221. Ha J, Park C, Park S. PMAMCA: prediction of microRNA-disease association utilizing a matrix completion approach. *BMC Syst Biol* 2019;**13**:33.
222. Chen X, Zhu CC, Yin J. Ensemble of decision tree reveals potential miRNA-disease associations. *PLoS Comput Biol* 2019;**15**:e1007209.
223. Chen Z, Wang X, Gao P, *et al*. Predicting disease related microRNA based on similarity and topology. *Cell* 2019;**8**:1405.
224. Chen X, Wang CC, Yin J, *et al*. Novel human miRNA-disease association inference based on random forest. *Mol Ther Nucleic Acids* 2018;**13**:568–79.
225. Fu L, Peng Q. A deep ensemble model to predict miRNA-disease association. *Sci Rep* 2017;**7**:14482.
226. Jiang Q, Wang G, Jin S, *et al*. Predicting human microRNA-disease associations based on support vector machine. *Int J Data Min Bioinform* 2013;**8**:282–93.
227. Dong Y, Sun Y, Qin C, *et al*. EPMDA: edge perturbation based method for miRNA-disease association prediction. *IEEE/ACM Trans Comput Biol Bioinform* 2020;**17**:2170–5.
228. Le D, Pham V, Nguyen TT An ensemble learning-based method for prediction of novel disease-microRNA associations. In: *2017 9th International Conference on Knowledge and Systems Engineering (KSE)*. Hue, VIETNAM: IEEE, 345 E 47TH ST, NEW YORK, NY 10017 USA, 2017.
229. Ji B, You Z, Cheng L, *et al*. Predicting miRNA-disease association from heterogeneous information network with GraRep embedding model. *Sci Rep* 2020;**10**:6658.
230. Jin L, Sai Z, Tao L, *et al*. Neural inductive matrix completion with graph convolutional networks for miRNA-disease association prediction. *Bioinformatics* 2020;**36**:2538–46.
231. Li C, Liu H, Hu Q, *et al*. A novel computational model for predicting microRNA–disease associations based on heterogeneous graph convolutional networks. *Cell* 2019;**8**:977.
232. Pan X, Shen H. Scoring disease-microRNA associations by integrating disease hierarchy into graph convolutional networks. *Pattern Recogn* 2020;**105**:107385.
233. Pan X, Shen H. Inferring disease-associated MicroRNAs using semi-supervised multi-label graph convolutional networks. *iScience* 2019;**20**:265–77.
234. Yang L, Qiu C, Jian T, *et al*. HMDD v2.0: a database for experimentally supported human microRNA and disease associations. *Nucleic Acids Res* 2014;**42**:D1070–4.
235. Huang Z, Shi J, Gao Y, *et al*. HMDD v3. 0: a database for experimentally supported human microRNA–disease associations. *Nucleic Acids Res* 2019;**47**:D1013–7.
236. Wang D, Wang J, Lu M, *et al*. Inferring the human microRNA functional similarity and functional network based on microRNA-associated diseases. *Bioinformatics* 2010;**26**:1644–50.
237. Wang X, Zhu M, Bo D, *et al*. AM-GCN: adaptive multi-channel graph convolutional networks. In: *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge*

Discovery & Data Mining. San Diego, CA, USA: ACM, New York, NY, USA, 2020, pp. 1243–53.
238. Demšar J. Statistical comparisons of classifiers over multiple data sets. *J Mach Learn Res* 2006;**7**:1–30.
239. Friedman M. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Publ Am Stat Assoc* 1939;**32**:675–701.
240. Friedman M. A comparison of alternative tests of significance for the problem of *m* rankings. *Ann Math Stat* 1940;**11**:86–92.
241. Zhu R, Ji C, Wang Y, *et al*. Heterogeneous graph convolutional networks and matrix completion for miRNA-disease association prediction. *Front Bioeng Biotechnol* 2020;**8**:901.
242. Jiang Q, Wang Y, Hao Y, *et al*. miR2Disease: a manually curated database for microRNA deregulation in human disease. *Nucleic Acids Res* 2009;**37**:D98–104.
243. Yang Z, Ren F, Liu C, *et al*. dbDEMC: a database of differentially expressed miRNAs in human cancers. *BMC Genomics* 2010;**11**(Suppl 4):S5.
244. Cho WCS, Chow ASC, Au JSK. MiR-145 inhibits cell proliferation of human lung adenocarcinoma by targeting EGFR and NUDT1. *RNA Biol* 2011;**8**:125–31.
245. Grose D, Morrison DS, Devereux G, *et al*. The impact of comorbidity upon determinants of outcome in patients with lung cancer. *Lung Cancer* 2015;**87**:186–92.
246. Zhang H, Zhang H, Zhao M, *et al*. MiR-138 inhibits tumor growth through repression of EZH2 in non-Small cell lung cancer. *Cell Physiol Biochem* 2013;**31**:56–65.
247. Goh JN, Loo SY, Datta A, *et al*. microRNAs in breast cancer: regulatory roles governing the hallmarks of cancer. *Biol Rev* 2016;**91**:409–28.
248. Foley NH, O'Neill LA. miR-107: a toll-like receptor-regulated miRNA dysregulated in obesity and type II diabetes. *J Leukoc Biol* 2012;**92**:521–7.
249. Mishra R, Singh SK. HIV-1 tat C modulates expression of miRNA-101 to suppress VE-cadherin in human brain microvascular endothelial cells. *J Neurosci* 2013;**33**:5992–6000.
250. Polikepahad S, Knight JM, Naghavi AO, *et al*. Proinflammatory role for let-7 microRNAS in experimental asthma. *J Biol Chem* 2010;**285**:30139–49.
251. Jeong HC, Kim EK, Lee JH, *et al*. Aberrant expression of let-7a miRNA in the blood of non-small cell lung cancer patients. *Mol Med Rep* 2011;**4**:383–7.
252. Earle JSL, Luthra R, Romans A, *et al*. Association of microRNA expression with microsatellite instability status in colorectal adenocarcinoma. *J Mol Diagn* 2010;**12**:433–40.
253. Müller DW, Bosserhoff A. Integrin beta 3 expression is regulated by let-7a miRNA in malignant melanoma. *Oncogene* 2008;**27**:6698–706.
254. Quesne JL, Jones J, Warren J, *et al*. Biological and prognostic associations of miR-205 and let-7b in breast cancer revealed by in situ hybridization analysis of micro-RNA expression in arrays of archival tumour tissue. *J Pathol* 2012;**227**:306–14.
255. Fazio PD, Montalbano R, Neureiter D, *et al*. Downregulation of HMGA2 by the pan-deacetylase inhibitor panobinostat is dependent on hsa-let-7b expression in liver cancer cell lines. *Exp Cell Res* 2012;**318**:1832–43.
256. Williams AE. Functional aspects of animal microRNAs. *Cell Mol Life Sci* 2008;**65**:545–62.
257. Wei W, Hu Z, Fu H, *et al*. MicroRNA-1 and microRNA-499 downregulate the expression of the ets1 proto-oncogene in HepG2 cells. *Oncol Rep* 2012;**28**:701–6.
258. Kojima S, Chiyomaru T, Kawakami K, *et al*. Tumour suppressors miR-1 and miR-133a target the oncogenic function of purine nucleoside phosphorylase (PNP) in prostate cancer. *Br J Cancer* 2012;**106**:405–13.
259. Pavicic W, Perkiö E, Kaur S, *et al*. Altered methylation at MicroRNA-associated CpG Islands in hereditary and sporadic carcinomas: a methylation-specific multiplex ligation-dependent probe amplification (MS-MLPA)-based approach. *Mol Med* 2011;**17**:726–35.
260. Suzuki H, Takatsuka S, Akashi H, *et al*. Genome-wide profiling of chromatin signatures reveals epigenetic regulation of MicroRNA genes in colorectal cancer. *Cancer Res* 2011;**71**:5646–58.
261. Migliore C, Martin V, Leoni VP, *et al*. MiR-1 downregulation cooperates with MACC1 in promoting MET overexpression in human colon cancer. *Clin Cancer Res* 2012;**18**:737–47.