Journal Title Here, 2024, pp. 1–10

doi: DOI HERE
Advance Access Publication Date: Day Month Year
Paper

PAPER

SMCLMDA: statistical meta-paths contrastive learning for predicting miRNA-disease multidimensional relationships

Yangfeng Zhu,[1] Yongxian Fan,[1,*] Guicong Sun[1] and Xueping Li[1]

[1]School of Computer Science and Information Security, Guilin University of Electronic Technology, 541004, Guilin, China
[*]Corresponding author. yongxian.fan@gmail.com

Abstract

The accurate identification of miRNA-disease associations plays a crucial role in biomedical research and clinical applications. However, most research focuses on the existence of the association, without conducting further exploration. In this study, we propose a novel statistical meta-path contrastive learning-based approach (SMCLMDA), which aims to accurately identify the multidimensional relationships(up/down-regulation and causal/non-causal) between miRNAs and diseases. SMCLMDA first obtains the local structural information of miRNAs and diseases based on Node2Vec and uses it as the initial node input of the graph convolutional neural network (GCN). Then, the meta-path view constructed by the statistical method further enhances the representation of the similarity view via a contrastive learning strategy. Finally, SMCLMDA calculates the predicted probability of the multidimensional relationships of miRNA-disease via a multilayer perceptron(MLP). The experimental results demonstrate that SMCLMDA surpasses current state-of-the-art computational methods in predicting traditional associations, as well as up/down-regulation and causality/non-causality relationships. The case study additionally validates the effectiveness of SMCLMDA in identifying potential miRNA-disease associations.

Key words: statistical meta-path, multidimensional relationships, contrastive learning, up/down-regulation, causal/non-causal

1 Introduction

MicroRNAs (miRNAs) are a subset of non-coding RNAs, usually an endogenous short RNA molecule approximately 22 nucleotides in length [1]. miRNAs participate in numerous significant biological processes, including cell proliferation, differentiation, metabolism, and apoptosis, by repressing mRNA expression in organisms [2]. Additionally, research indicates that oncogenic miR-21-5p is elevated in rectal, breast, and lung cancers [3]. Therefore, the exploration of the multidimensional relationships between miRNAs and diseases is conducive to the development of disease research from single-factor analysis to systems biology and precision medicine.

In recent years, several standardized databases have been developed based on published literature, such as HMDD v4.0 [4], miR2Disease [5], and miRCancer [6]. HMDD v4.0 is an experimentally supported and validated database of classical miRNA-disease associations, with two additional categories: exosomal miRNAs and virus-encoded miRNAs. miR2Disease is a database curated manually, which is designed to provide a comprehensive resource on miRNA dysregulation across diverse human diseases. miRCancer specifically documents the associations between miRNAs and cancers. These public databases offer valuable references for studying miRNA-disease relationships.

To date, relatively scant information has been identified regarding miRNA-disease associations, especially causality/non-causality relationships. Additionally, most miRNA-disease associations originate from biological wet experiments, which are costly and time-consuming [7]. Conversely, the utilization of computational approaches to uncover potential miRNA-disease relationships is both efficient in time and labor-saving. Typically, these computational approaches can be generally categorized into three groups: similarity-based, machine learning, and graph-based methods. Similarity-based approaches prioritize disease-associated miRNAs, assuming that miRNAs with functional similarities are often linked to diseases with comparable phenotypes [8, 9]. More typical similarity calculation methods in the prediction of miRNA-disease associations include disease semantic similarity [10], miRNA functional similarity [11], and Gaussian interaction profile (GIP) kernel similarity [12]. For example, Han et al. [13] calculated the multiple biometrics matrices and obtained the prediction results of each similarity matrix based on graph semi-supervised learning. Wang et al. [14] proposed a nonlinear fusion strategy to mitigate the noise resulting from similarity differences among different data sources and

© The Author 2024. Published by Oxford University Press. All rights reserved. For permissions, please e-mail: journals.permissions@oup.com

their influence on network quality. Yu et al. [15] constructed multiple biological meta-paths based on genes and combined the characteristics of nodes with the information on the meta-paths to calculate the association scores. Wang et al. [16] built a multi-layer heterogeneous network combining miRNA, disease, and gene similarities to predict potential associations through functionally similar genes. Similarity-based approaches, while making full use of the validated associations, depend heavily on the selection of similarity metrics and the quality of feature representation. In addition, these methods are challenging to extend to multimodal data, especially when multiple similarities or data types are considered simultaneously.

Machine learning technology selects suitable classifiers for prediction via data processing and feature selection. For instance, Liu et al. [17] based on validated associations by first using an autoencoder to obtain potential features, and then combining similarities to obtain predicted scores via XGBoost. In addition, the extreme imbalance of data can lead to the prediction model being biased toward the majority class, thereby influencing the prediction accuracy. Thus, Dai et al. [18] used a resampling technique to tackle the sample imbalance and predicted miRNA-disease associations through a soft voting approach in integrated learning. Wang et al. [19] initially trained the autoencoder through an unsupervised approach and then optimized the model in a supervised way with balanced samples. However, machine learning-based methods are highly data-dependent, and the sparse nature of miRNA-disease association matrices heightens the risk of overfitting. In recent years, as the demand for complex biological network analysis increases, to address the challenges of high-dimensional and sparse data, graph-based approaches have demonstrated strong performance. Graph Convolutional Neural Networks [20] are particularly well-suited for non-Euclidean data structures, capturing complex relationships between neighboring nodes across different orders and learning node embeddings automatically. For example, Peng et al. [21] constructed multi-hypergraphs based on the similarity and obtained consistent higher-order representations of miRNAs and diseases in each view via hypergraph convolution. Ning et al. [22] introduced virtual nodes to create heterogeneous miRNA-disease hypergraphs, enriching the link information. Lou et al. [23] established a multimodal network based on the miRNA(disease) similarities and verified associations, leveraging GCN to aggregate muti-order neighbor information within the network. Tang et al. [24] obtained embedded representations of multiple views via GCN and integrated the multi-view information through the pipeline attention mechanism. He et al. [25] built miRNA-gene and disease-gene sub-networks by gene to alleviate the data sparsity problem. Although graph-based methods are powerful, the issue of excessive smoothing in GCN cannot be overlooked. In addition, graph data frequently contains noise, and incorrect edges or node connections may mislead the model, thereby degrading the model's performance.

Although existing approaches have achieved good results, most of them only obtain information from a single similarity perspective. Additionally, some computational methods enrich link information by introducing biological knowledge, such as introducing genes as intermediaries to form a multi-layer heterogeneous network. While this strategy adds much biological a priori knowledge, it also introduces more noise and computational complexity. miRNA-disease relationships are complex and diverse, encompassing multiple regulatory mechanisms. For example, with validated associations, these relationships can be further subdivided into causal/non-causal and up/down-regulation. Among them, causal means that miRNAs directly affect the occurrence of disease, and non-causal means that miRNAs are related to disease but are not the direct cause of disease. Up-regulation and down-regulation indicate that miRNAs promote or inhibit disease development, respectively. Most studies have primarily explored the duality of miRNA and disease, and have not yet conducted an in-depth analysis.

Thus, we propose a novel computational approach based on statistical meta-path contrastive learning (SMCLMDA) to fully explore the traditional binary associations, up/down-regulation, and causal/non-causal relationships. SMCLMDA enriches the information of similarity views by constructing additional meta-path views and utilizes the contrastive learning strategy. Additionally, to have certain graph structure knowledge at the beginning of model training, the Node2Vec is utilized to obtain local structural information of miRNAs and diseases as the initial node features of GCN. The SMCLMDA model framework is shown in Fig 1.

2 Materials and methods

2.1 Datasets

In this work, we collated datasets of various relationship types for experiments. Among these, HMDD v4.0 [4]contains 1817 miRNAs, 2360 diseases, and 53,530 records of experimentally supported miRNA-disease associations. miR2Disease [5] comprises 349 miRNAs, 163 diseases, and 3,273 known associations. Notably, miR2Disease adds additional information on the up-or down-regulation of certain miRNA-disease relationships beyond the known associations. miRTarBase [26] focuses on experimentally validated miRNA target genes, ensuring that all included miRNAs have been experimentally confirmed. The Disease Ontology (DO) database [27] offers a comprehensive and standardized classification of diseases, each of which has a unique Disease Ontology ID (DOID). We downloaded the Medical Subject Headings (MeSH) 2024 [28] for data preprocessing. In this paper, we focused on studying human diseases. During the organization of the HMDD v4.0 dataset, we manually removed diseases that were not categorized under MeSH category C or lacked a DOID identifier, as well as miRNAs not included in miRTarBase. The same data filtering process was applied to the subsequent datasets.

Zhang et al. [29] explored the up/down-regulation relationships between miRNAs and diseases. Given the limited number of such relationships in miR2Disease, we integrated its data with the dataset from Zhang et al.'s study. When studying up/down-regulation relationships in the dataset miR2Disease, we removed the records that had associations in miR2Disease but lacked labeled up/down-regulation. For causality/non-causality prediction, we used the dataset yanzheng organized by Han et al. [13]. Additionally, Han et al. [13] also curated the HMDD v3.2 dataset for predicting traditional associations. We conducted secondary filtering on these datasets. It is important to note that when removing diseases that are not in MeSH, we considered different representations of the same disease. For example, "hepatocellular carcinoma" is referred to as "hepatocellular carcinoma (HCC)" in the miR2Disease database, whereas it appears as "Carcinoma, Hepatocellular" in MeSH. To facilitate the subsequent calculation of disease

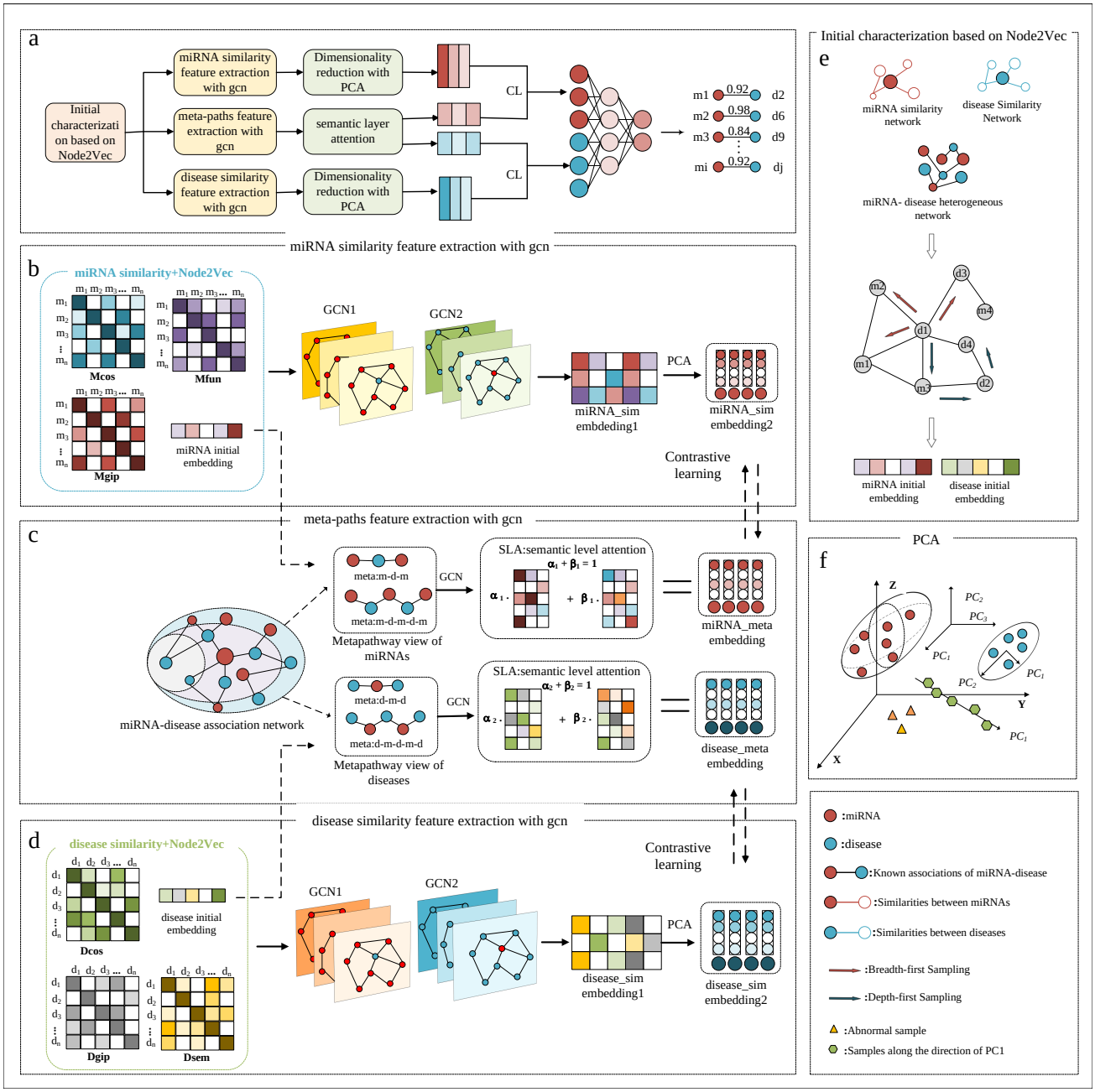

Fig. 1. (a) Overview of the SMCLMDA framework. (b) Higher-order representations of the miRNA similarity views are learned through the GCN module. (c) Higher-order representations of different meta-path views of miRNAs and diseases are captured via the GCN module, and information from different views is fused based on the SLA strategy. (d) Higher-order representations of disease similarity views are learned through the GCN module. (e) Local structural information of miRNAs and diseases is obtained via Node2Vec as initial features of nodes for GCN in (b), (c), and (d). (f) PCA dimensionality reduction strategy is utilized in modules (b) and (d).

semantic, the disease name needs to be kept consistent with the MeSH. Table 1 presents the details of the final dataset.

2.2 Similarity view construction

In this experiment, we define the known association matrix as $A \in \mathbb{R}^{n_m \times n_d}$, where n_m and n_d represent the number of miRNA and disease respectively. miRNA similarity views include cosine similarity M_{cos}, GIP kernel similarity M_{gip}, and functional similarity M_{fun}. Disease similarity views include cosine similarity D_{cos}, GIP core similarity D_{gip}, and semantic similarity D_{sem}. For the detailed calculation process, please refer to the Supplementary File(Section 1-4).

2.3 Construction of the meta-path view

Meta-path is the path of connections formed by different edge types within a heterogeneous network. Through the use of meta-path, potential associations between different objects formed by the interactions of multiple edge types can be identified [30, 31]. Given the heterogeneity of the miRNA-disease interaction network, we used the meta-path to capture

Table 1. Overview of datasets utilized in this study.

Dataset	miRNA	disease	association
HMDD v4.0(0/1)	728	884	13511
miR2Disease + zhang(0/1)	270	246	3534
miR2Disease + zhang(1/-1)	270	246	2850
HMDD v3.2(0/1)	533	293	4205
yanzheng(causal/no-causal)	503	277	8158

where association (2850) contains 1507 up-regulation and 1343 down-regulation, and association (8158) contains 1986 causal and 6172 noncausal

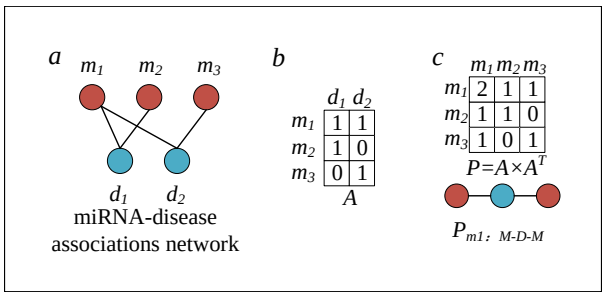

Fig. 2. Example of generating the meta-path P_{m_1}. (a) miRNA-disease known association network. (b) miRNA-disease association matrix. (c) construction of the meta-path MDM through matrix multiplication.

the intricate relationships between nodes. Specifically, a meta-path is expressed as $N_1 \xrightarrow{R_1} N_2 \xrightarrow{R_2} \cdots \xrightarrow{R_l} N_l$, where the complex relationship between node N_1 and N_l is represented as $R = R_1 \circ R_2 \circ \cdots \circ R_l$, with $\circ$ denoting the relational conformity operator.In this paper, two statistically significant meta-paths, namely, $P_{m1} = MDM$, $P_{m2} = MDMDM$, $P_{d1} = DMD$ and $P_{d2} = DMDMD$, were constructed for miRNA and disease respectively, and were calculated as follows:

$$P_{m1} = A \times A^T \tag{1}$$

$$P_{m2} = P_{m1}^2 = (A \times A^T)^2 \tag{2}$$

$$P_{d1} = A^T \times A \tag{3}$$

$$P_{d2} = (P_{d1})^2 = (A^T \times A)^2 \tag{4}$$

The specific construction of the meta-path is illustrated in Fig 2. Each off-diagonal element P_{ik} in the meta-path matrix , illustrated in Fig. 2c, indicates the number of distinct types of the meta-path between miRNA m_i and m_k. For example, $P_{13} = 1$ signifies that only one type of the meta-path exists between miRNA m_1 and m_3, indicating that m_1 and m_3 are directly connected via disease d_2. This connection is represented in either Fig. 2a or Fig. 2b. The diagonal element $P_{11} = 2$ indicates that m_1 is associated with two different diseases.

Ultimately, we constructed meta-path views for miRNA $VP_{m1} \in \mathbb{R}^{n_m \times n_m}$ and $VP_{m2} \in \mathbb{R}^{n_m \times n_m}$, as well as for disease, where $VP_{d1} \in \mathbb{R}^{n_d \times n_d}$ and $VP_{d2} \in \mathbb{R}^{n_d \times n_d}$.

2.4 Node2Vec encoding strategy

In previous studies, the node features in GCN were often initialized randomly. Such an approach may lead to initial representations that lack sufficient information, failing to obtain the structural and semantic characteristics of miRNAs and diseases. In this study, we employ Node2Vec to generate semantically enriched feature vectors, which serve as the initial node feature input for GCN. For detailed process, please refer to the Supplementary File(Section 5). Finally, we obtain the initial feature embedding representations of miRNAs and diseases, represented as $mi_emb \in \mathbb{R}^{n_m \times n_g}$ and $di_emb \in \mathbb{R}^{n_d \times n_g}$ respectively. In this paper, the embedding size n_g is set to 64.

2.5 Learning of node embedding in similarity network

We construct multiple networks for miRNA-miRNA and disease-disease interactions by using the similarities computed in subsection 2.2 as adjacency matrices. The set of miRNAs networks, denoted as S_m, and the set of diseases networks, denoted as S_d, are defined as follows:

$$S_m = \{M_{cos}, M_{gip}, M_{fun}\} \tag{5}$$

$$S_d = \{D_{cos}, D_{gip}, D_{sem}\} \tag{6}$$

Recently, graph convolutional neural networks have garnered significant attention for their robust feature aggregation capabilities. GCN effectively captures dependencies among nodes and learns the higher-order feature representations of individual nodes. We apply vanilla GCN [20] to the constructed similarity matrix to obtain the embedding representations of nodes. Following the information aggregation and update rules of GCN, we calculate miRNA node embedding as follows:

$$H_m^{(l+1)} = \sigma\left(\tilde{D}_m^{-\frac{1}{2}} \tilde{S}_m \tilde{D}_m^{-\frac{1}{2}} H_m^{(l)} W_m^{(l)}\right) \tag{7}$$

where $\sigma(\cdot)$ is the activation function. $\tilde{S}_m = S_m + I$ and I represents the identify matrix to retain the original node information. $\tilde{D}_m$ denotes the degree matrix of $\tilde{S}_m$. $H_m^{(l)}$ refers to the embedding of the node at the l-th, and $W_m^{(l)} \in \mathbb{R}^{n_g \times n_g}$ is the learnable weight matrix.

The embeddings of different views of miRNA, obtained after GCN processing, are represented as follows:

$$H_{m_{cos}} = \left\{H_{m_{cos}}^{(1)}, H_{m_{cos}}^{(2)}, \cdots, H_{m_{cos}}^{(l)}\right\} \tag{8}$$

$$H_{m_{gip}} = \left\{H_{m_{gip}}^{(1)}, H_{m_{gip}}^{(2)}, \cdots, H_{m_{gip}}^{(l)}\right\} \tag{9}$$

$$H_{m_{fun}} = \left\{H_{m_{fun}}^{(1)}, H_{m_{fun}}^{(2)}, \cdots, H_{m_{fun}}^{(l)}\right\} \tag{10}$$

The embedding representations obtained from each view were concatenated, and then the PCA technique was applied for dimensionality reduction to obtain the final miRNA embedding representation, denotes as $X_m \in \mathbb{R}^{n_m \times n_g}$, as shown below:

$$X_m = PCA\left(cat(H_{m_{cos}}, H_{m_{gip}}, H_{m_{fun}})\right) \tag{11}$$

Similarly, the final embedding of the disease is represented as $Y_d \in \mathbb{R}^{n_d \times n_g}$

2.6 Learning of node embedding in meta-path network

According to the previously calculated miRNA and disease meta-path views VP_{m1}, VP_{m2}, VP_{d1}, and VP_{d2}. We employ the vanilla GCN to obtain the embedded features of miRNAs and diseases. The miRNA embedding representation is calculated as follows:

$$H_{VP_{mk}}^{(l+1)} = \sigma\left(\tilde{D}_{VP_{mk}}^{-\frac{1}{2}}\tilde{Q}_{VP_{mk}}\tilde{D}_{VP_{mk}}^{-\frac{1}{2}}H_{VP_{mk}}^{(l)}W_{VP_{mk}}^{(l)}\right), k \in \{1,2\} \tag{12}$$

where $\tilde{Q} = Q_{VP_{mk}} + I$, with $Q_{VP_{mk}} = \{VP_{m1}, VP_{m2}\}$. $H_{VP_{mk}}^{(l)}$ represents the node embedding at the l-th layer.

The embeddings of miRNA meta-path views across different graph convolutional layers are represented as follows:

$$\left\{H_{VP_{m1}}^{(1)}, H_{VP_{m1}}^{(2)}, \cdots, H_{VP_{m1}}^{(l)}, H_{VP_{m2}}^{(1)}, H_{VP_{m2}}^{(2)}, \cdots, H_{VP_{m2}}^{(l)}\right\} \tag{13}$$

Inspired by Zhen et al. [31], we use SLA to process Eq.(13). The final miRNA embedding, $P_m \in \mathbb{R}^{n_m \times n_g}$, is computed as follows:

$$P_m = \sum_{i=1}^{l} \alpha_i H_{VP_{mk}}^{(i)}, k \in \{1,2\} \tag{14}$$

$$\alpha_i = \frac{exp\left(a^T \cdot tanh(W \cdot H_{VP_{mk}}^{(i)} + b)\right)}{\sum_{j=1}^{l} exp\left(a^T \cdot tanh(W \cdot H_{VP_{mk}}^{(j)} + b)\right)} \tag{15}$$

where $W \in \mathbb{R}^{n_g \times n_g}$ and $b \in \mathbb{R}^{n_g \times 1}$ are learnable weight paramenters, and $a \in \mathbb{R}^{n_g \times 1}$ denotes the attention weight vector.

A comparable approach is employed to drive the final embedding of the disease across various meta-path views, denoted as $P_d \in \mathbb{R}^{n_d \times n_g}$.

2.7 Contrastive learning strategy

The two view embeddings obtained for miRNAs, X_m and P_m, were projected into the same latent space using a linear layer to ensure that subsequent similarity calculations are conducted in the same potential space:

$$V_{m_i}^{sim} = W_{sim}X_m + b_{sim} \tag{16}$$

$$V_{m_i}^{mp} = W_{mp}P_m + b_{mp} \tag{17}$$

where $W_{sim} \in \mathbb{R}^{n_g \times n_g}$ and $W_{mp} \in \mathbb{R}^{n_g \times n_g}$ denote the learnable weight matrices, while $b_{sim} \in \mathbb{R}^{n_g \times 1}$ and $b_{mp} \in \mathbb{R}^{n_g \times 1}$ represent the bias vectors.

When screening positive sample pairs, the traditional method of using the same miRNAs under different views as positive sample pairs may fail to capture complex semantic relationships and reveal potential miRNA-disease associations. We consider that the richer the variety of meta-paths between two different miRNAs indicates higher similarity, allowing them to be regarded as positive sample pairs. Assuming that there are two distinct views G_1 and G_2, where m_i and $m_i^{'}$ represent instances of the miRNA in G_1 and G_2 respectively, the $\left(m_i, m_i^{'}\right)$ is considered a positive sample pairs. For different instances, as depicted in the matrix P in Fig. 2c, if there is no common disease between m_3 and m_2, but m_3 and m_1 are connected through a disease, then $\left(m_3, m_1^{'}\right)$ or $\left(m_3^{'}, m_1\right)$ are classified as positive sample pairs, whereas $\left(m_3, m_2^{'}\right)$ or $\left(m_3^{'}, m_2\right)$ are categorized as negative samples. Finally, all positive pairs are denoted as P_{m_i}, and all negative pairs are represented as N_{m_i}.

Based on the previously defined P_{m_i} and N_{m_i}, the contrastive loss for the miRNA similarity view is formulated as follows:

$$loss_{m_i}^{sim} = -log\frac{\sum_{m_j \in P_{m_i}} exp\left(\frac{sim\left(V_{m_i}^{sim}, V_{m_j}^{mp}\right)}{\tau}\right)}{\sum_{m_k \in P_{m_i} \cup N_{m_i}} exp\left(\frac{sim\left(V_{m_i}^{sim}, V_{m_k}^{mp}\right)}{\tau}\right)} \tag{18}$$

where $sim\,(\cdot)$ represents the cosine similarity between miRNAs m_i and m_j, and τ denotes the temperature coefficient.

Similarly, the contrastive loss for the miRNA meta-path view is formulated as follows:

$$loss_{m_i}^{mp} = -log\frac{\sum_{m_j \in P_{m_i}} exp\left(\frac{sim\left(V_{m_i}^{mp}, V_{m_j}^{sim}\right)}{\tau}\right)}{\sum_{m_k \in P_{m_i} \cup N_{m_i}} exp\left(\frac{sim\left(V_{m_i}^{mp}, V_{m_k}^{sim}\right)}{\tau}\right)} \tag{19}$$

The total contrastive loss of miRNA embedding representation is expressed as follows:

$$loss_{mi} = \lambda_m loss_{m_i}^{sim} + (1 - \lambda_m)\, loss_{m_i}^{mp} \tag{20}$$

where λ_m represents the contribution factor used to balance the similarity view and the meta-path view.

Similarly, the total contrastive loss of the disease embedding representation is expressed as follows:

$$loss_{dj} = \lambda_d loss_{d_j}^{sim} + (1 - \lambda_d)\, loss_{d_j}^{mp} \tag{21}$$

SMCLMDA learns the node higher-order representations through backpropagation optimization. Finally, the embedding of miRNA (X_m) and disease (Y_d) are normalized with layerNorm to obtain $\tilde{X}_m$ and $\tilde{Y}_d$.

2.8 Probabilistic forecasting

In this study, The MLP is employed to predict the multidimensional relationships between miRNAs and diseases. First, the Hadamard product of $\tilde{X}_m$ and $\tilde{Y}_d$ is calculated and input into the MLP to generate the association score y_{score}. SMCLMDA employs the binary cross-entropy as its loss function, which is formulated as follows:

$$loss_{pre} = \sum_{(i,j) \in y^+ \cup y^-} y_{ij} log y_{score} + (1 - y_{ij})\, log\,(1 - y_{score}) \tag{22}$$

where (i, j) represents miRNA-disease pairs, while y^+ and y^- correspond to positive and negative sample sets respectively. For $(i, j) \in y^+$, $y_{ij} = 1$; otherwise, $y_{ij} = 0$.

Combined with the contrastive losses in subsection 2.7, the overall loss is expressed as follows:

$$loss = C\,(loss_{mi} + loss_{dj}) + (1 - C)\, loss_{pre} \tag{23}$$

where C is the contribution factor balancing the contrastive and prediction lossess.

3 Experiment and result

3.1 Experimental setup

To comprehensively assess the SMCLMDA's performance in predicting miRNA-disease multidimensional relationships, the benchmark dataset described in Section 2.1 was divided using

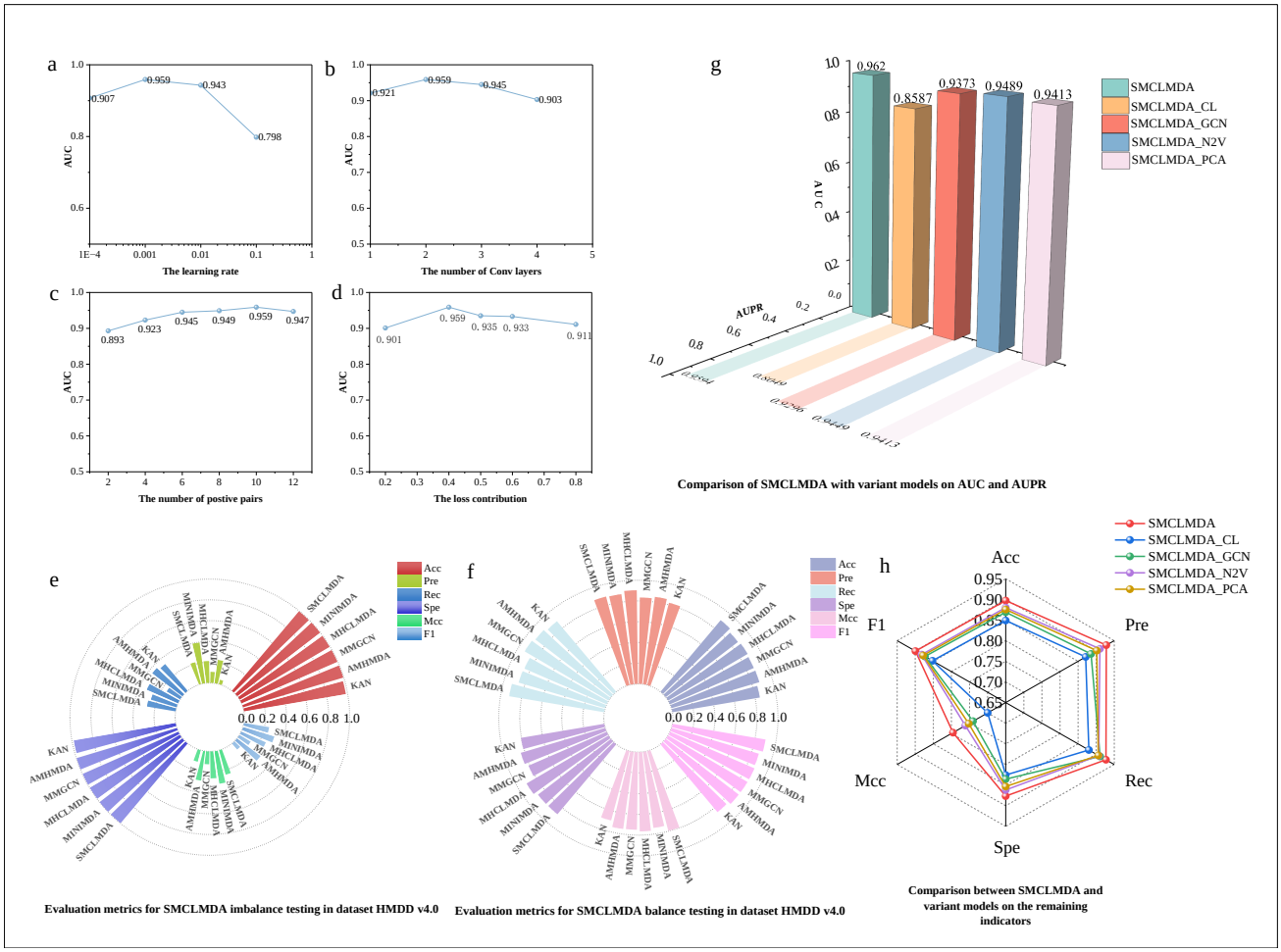

Fig. 3. Analysis of SMCLMDA performance evaluation results. (a) The influence of different learning rates on model performance. (b) The impact of varying numbers of convolutional layers on model performance. (c) The influence of selecting different numbers of positive sample pairs on model performance. (d) The effect of the contrastive loss contribution on model performance. (e) The comparison of imbalance test performance between SMCLMDA and baseline approaches on the HMDD v4.0 (0/1) dataset. (f) The comparison of balanced test performance between SMCLMDA and the baseline method on the HMDD v4.0 (0/1) dataset. (g) The AUC and AUPR comparison between SMCLMDA and its variant models. (h) The comparison of SMCLMDA with variant models in terms of other metrics.

different strategies.For detailed process, please refer to the Supplementary File(Section 6).

The performance metrics employed to evaluate the SMCLMDA included the Area Under the Receiver Operating Characteristic Curve (AUC), the Area Under the Precision-Recall Curve (AUPR), Accuracy, Precision, Recall, Specificity, Mathew's Correlation Coefficient (MCC), and F1-score.

3.2 Parameter analysis

In this subsection, we analyzed how model parameters influence the performance of SMCLMDA. Specifically, we investigated the effects of different learning rates l_r, the count of convolutional layers L, the number of postive sample pairs used for constructing the meta-path S, and the contribution coefficients balancing the contrastive loss and prediction loss C on the prediction results. All experiments were performed on the HMDD v4.0 dataset using stratified sampling.

The learning rate significantly influences the convergence of the model. A larger learning rate facilitates faster convergence but may lead to suboptimal solutions, while a smaller learning rate enables the model to explore the loss function space more thoroughly at the cost of increased training time. As shown in Fig. 3a, the learning rate l_r was selected from the set $\{0.0001, 0.001, 0.01, 0.1\}$, and the model achieced the best performance with $l_r = 0.001$. The number of convolutional layers affects feature extraction and the model's generalization ability. Too few convolutional layers may not be able to extract the deep features, introducing excessive bias. Conversely, an increased number of layers heightens the risk of overfitting while extracting complex features. As illustrated in Fig. 3b, the count of convolutional layers L was chosen from the set $\{1, 2, 3, 4\}$, with the model achieving optimal performance when $L = 2$. In contrastive learning, a small number of positive sample pairs may result in insufficient learning signals and reduce the ability of the model to distinguish between similar samples. The number of positive sample logarithms S was selected from the set $\{2, 4, 6, 8, 10, 12\}$. As depicted in Fig. 3c, SMCLMDA achieved the best performance with $S = 10$. Finally, to balance the contributions of contrastive loss and prediction loss, the contrastive loss contribution C was varied (where $1 - C$ represents the contribution of prediction loss) across the set $\{0.2, 0.4, 0.5, 0.6, 0.8\}$. As shown in Fig.

Table 2. Performance comparison of all methods in predicting traditional associations on unbalanced test sets.

Methods	HMDD v4.0(0/1)		miR2Disease + zhang(0/1)		HMDD v3.2(0/1)	
	AUC	AUPR	AUC	AUPR	AUC	AUPR
KAN	0.9177	0.0282	0.8652	0.1102	0.9131	0.0667
AMHMDA	0.9338	0.1746	0.8850	0.0972	0.9257	0.1169
MMGCNA	0.9343	0.0705	0.8716	**0.1194**	0.9132	0.0999
MHCLMDA	0.9223	0.1933	0.8556	0.0999	0.9178	0.1486
MINIMDA	0.9404	**0.2419**	0.8793	0.1069	0.9250	**0.1512**
SMCLMDA	**0.9618**	0.1460	**0.9020**	0.0882	**0.9362**	0.0861

3d, SMCLMDA achieved the highest prediction scores when $C = 0.4$. SMCLMDA employed the Adam optimizer, and detailed hyperparameter combinations are provided in the Supplementary File(Section 7).

3.3 Comparison of SMCLMDA and other methods

To evaluate the superiority and generalization capability of SMCLMDA, it was compared against five state-of-the-art approaches: KAN [32], AMHMDA [22], MMGCN [24], MHCLMDA [21], and MINIMDA [23].

Additionally, to ensure an objective and fair comparison, the same dataset and miRNA and disease similarities were used for all methods. The parameter settings for SMCLMDA and each baseline method are provided in the Supplementary File(Section 7). For traditional association prediction, the corresponding results on the HMDD v4.0 (0/1), miR2Disease + zhang (0/1), and HMDD v3.2 (0/1) datasets are shown in Tables 2 and 3. Specifically, Table 2 summarizes the performance comparison between SMCLMDA and other approaches on the unbalanced test set, while Table 3 presents the corresponding performance on the balanced test set. The results indicate that SMCLMDA outperforms other methods in performance across both balanced and unbalanced test sets. Notably, in the balanced tests across the three datasets, SMCLMDA achieves improvements of 3%, 2%, and 1% in the AUC metric compared to the next best method respectively. Similarly, in the imblance test, SMCLMDA outperforms the next best method by 2%, 1.5%, and 1% in the AUC metric respectively. Fig. 3e and 3f illustrate the performance comparison of SMCLMDA with other baseline methods across the additional metrics in both balanced and unbalanced tests on the HMDD v4.0 (0/1) dataset. The detailed experimental results, including Accuracy and Recall, for the HMDD v4.0 (0/1), miR2Disease + zhang (0/1), and HMDD v3.2 (0/1) datasets are provided in the Supplementary File (Section 8.Tables 1, 2, and 3).

The results of SMCLMDA predicting up/down-regulation and causal/non-causal are shown in Table 4 and Table 5, respectively. For predicting the up/down-regulation relationship, SMCLMDA attained an AUC of 0.9105, which was 1.5% outperforming than the suboptimal method. Although MMGCN attains the highest Recall score of 0.9535, SMCLMDA outperforms the other methods across the remaining evaluation metrics. For predicting causal/non-causal relationship, SMCLMDA attains an AUC of 0.877, outperforming the next best method by 1.7%. Additionally, SMCLMDA outperforms other baseline in the other evaluation metrics, such as Accuracy and Recall. These results demonstrate that SMCLMDA effectively predicts the multi-dimensional relationships between miRNAs and diseases.

3.4 Ablation experiments

To validate the contribution of each component in SMCLMDA, four variants of the SMCLMDA model were designed for comparison. Specifically, ten-fold cross-validation was conducted on the HMDD v4.0 (0/1) dataset for performance evaluation. The variants are as follows: SMCLMDA_CL removes the contrastive learning loss. SMCLMDA_GCN replaces the GCN module with a linear layer. SMCLMDA_N2V replaces the initial embedding of nodes computed by Node2Vec using randomly initialized features. SMCLMDA_PCA removes the PCA dimensionality reduction step. The experimental results are shown in Fig. 3g, which demonstrates that the contrastive learning strategy contributes the most among the components of SMCLMDA. In the AUC evaluation metric, the contrastive learning strategy improves SMCLMDA's performance by approximately 10%, followed by GCN, which contributes an improvement of approximately 3%. Fig. 3h further compares SMCLMDA with the four variant models across other metrics, showing that SMCLMDA consistently outperforms all variants. In conclusion, our model further enhances the embedded representation of miRNAs and diseases based on GCN aggregation of neighbor information through a contrastive learning strategy, which ultimately improves the prediction performance of SMCLMDA. Detailed evaluation metrics for SMCLMDA and each variant model are provided in the Supplementary File (Section 8.Table 4).

3.5 SMCLMDA confidence analysis

To gain deeper insights into the model's convergence process and its ability to differentiate positive from negative samples during training, we conducted experiments on the HMDD v4.0 (0/1) dataset. As shown in Fig. 4a, when epoch=0, SMCLMDA exhibits limited ability to differentiate positive from negative samples, with the two categories being almost indistinguishable. As training progresses, the distribution of predicted scores gradually stabilizes. When epoch=400, the model has largely converged, with prediction scores for positive samples (label 1) concentrated in a higher range and those for negative samples (label 0) concentrated in a lower range. Typically, the prediction scores are indicative of the model's confidence, where values closer to 0 or 1 signify higher confidence in the prediction. Nonetheless, some undetected outliers persist, which represent samples that are difficult for the model to distinguish. Future research should further explore the characteristics of these outliers and their potential causes. As shown in Fig. 4b, the confusion matrix for the unbalanced test set reveals that while SMCLMDA performs well in predicting negative samples, its performance on positive samples requires further improvement. In contrast, as depicted in Fig. 4c, SMCLMDA demonstrates

Table 3. Performance comparison of all methods in predicting traditional associations on balanced test sets.

Methods	HMDD v4.0(0/1)		miR2Disease + zhang(0/1)		HMDD v3.2(0/1)	
	AUC	AUPR	AUC	AUPR	AUC	AUPR
KAN	0.9153	0.9202	0.8671	0.8738	0.9171	0.9117
AMHDA	0.9350	0.9297	0.8824	0.8755	0.9248	0.9272
MMGCN	0.9380	0.9333	0.8681	0.8767	0.9157	0.9171
MHCLMDA	0.9230	0.9425	0.8582	0.8630	0.9104	0.9254
MINIMDA	0.9378	0.9471	0.8704	0.8702	0.9169	0.9264
SMCLMDA	**0.9620**	**0.9593**	**0.9024**	**0.8899**	**0.9377**	**0.9338**

Table 4. Performance comparison of all methods in predicting up/down-regulation on the miR2Disease + zhang (1/-1) dataset.

Methods	Acc	Pre	Rec	Spe	Mcc	F1	AUC	AUPR
KAN	0.8380	0.8520	0.9442	0.6058	0.5888	0.8919	0.8687	0.9274
AMHMDA	0.8443	0.8594	0.9244	0.6696	0.6273	0.8906	0.8917	0.9422
MMGCN	0.8380	0.8345	**0.9535**	0.5855	0.6098	0.8898	0.8944	0.9467
MHCLMDA	0.7693	0.7785	0.9310	0.4159	0.4225	0.8472	0.7772	0.8598
MINIMDA	0.7961	0.8250	0.8945	0.5812	0.5099	0.8574	0.8320	0.9053
SMCLMDA	**0.8585**	**0.8732**	0.9297	**0.7029**	**0.6640**	**0.9001**	**0.9105**	**0.9545**

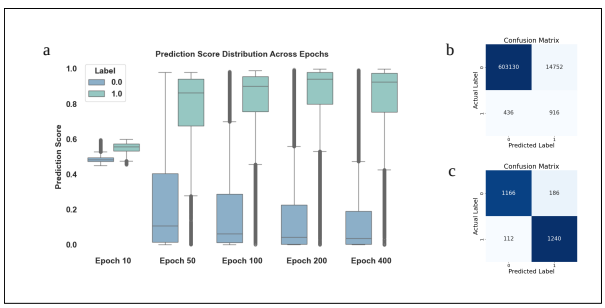

Fig. 4. Confidence analysis of SMCLMDA. (a) SMCLMDA discrimination between positive and negative samples. (b) Confusion matrix of SMCLMDA on the unbalanced test. (c) Confusion matrix of SMCLMDA on the balanced test.

fewer misclassified for both positive and negative samples on the balance test set. Overall, SMCLMDA achieves high confidence in predictions across both balanced and unbalanced data, further corroborating its superiority and robustness in predicting miRNA-disease multidimensional relationships.

4 Case studies

To further validate the capability of SMCLMDA to predict candidate miRNAs linked to potential diseases, case studies were conducted on lung, breast, and colorectal cancer. Initially, we trained the model using all validated associations from the HMDD v4.0 (0/1) dataset to generate a prediction score matrix. Subsequently, we screened the top 30 candidate miRNAs for each of the three cancers based on these prediction scores. These candidate miRNAs were then validated against the dbDEMC [33] and miRCancer [6] databases. The dbDEMC database provides different types of cancers and their associated miRNAs differential expression data, including 3,268 miRNAs and 40 cancer types. miRCancer integrates a large body of research literature and offers information on the association of miRNAs with a wide range of cancers, currently encompassing 57,984 miRNAs, 196 cancer types, and 9,080 associations.

The top 10 relevant candidate miRNAs for lung, breast, and colorectal cancers, as presented in Table 5(Supplementary File Section 8). The results show that all top 10 candidate miRNAs for the three cancers have been validated by the dbDEMC or miRCancer databases. For detailed results of the top 30 candidate miRNAs for each cancer type, please refer to the Supplementary File (Section 8.Tables 6, 7, and 8).

5 Conclsion

In this study, we developed a novel computational approach based on statistical meta-path contrastive learning (SMCLMDA), which was designed to accurately identify the traditional association, up/down-regulation, and causal/non-causal relationships. SMCLMDA enriches view information by constructing the statistical meta-path and enhances the representation of similar views based on contrastive learning. Additionally, Node2Vec is utilized to obtain the initial representation of nodes, so that the model has certain graph structure knowledge at the beginning of training, thereby accelerating the convergence of the model. Our method outperforms existing computational methods in predicting miRNA-disease associations, up/down-regulation, and causality/non-causality relationships, with particularly significant improvements in the prediction of causality/non-causality relationships. Cancer cases further validate the ability of SMCLMDA to predict potential association.

In future work, we aim to expand upon the known miRNA-disease associations by further collecting and organizing datasets related to up/down-regulation and causal/non-causal relationships. Currently, the meta-path view proposed in this paper is based on direct calculation by matrix multiplication in statistics, which lacks sufficient biological interpretability. To address this limitation, we will look for alternative views with biological to augment the information in the base similarity view. Additionally, we will consider the reliability of the information in the negative samples and screen the most representative negative samples for model training. Finally, to reduce the redundant information between views, we will try to fuse different views of miRNA and disease, enabling the model

Table 5. Performance comparison of all methods in predicting causal/non-causal on the yanzheng (causal/non-causal) dataset.

Methods	Acc	Pre	Rec	Spe	Mcc	F1	AUC	AUPR
KAN	0.8013	0.6607	0.8807	0.7590	0.6109	0.7550	0.8603	0.7559
AMHMDA	0.6178	0.4703	0.7842	0.5293	0.3022	0.5879	0.7306	0.5965
MMGCN	0.7680	0.6454	0.7387	0.7837	0.5082	0.6889	0.8508	0.7776
MHCLMDA	0.6855	0.4185	0.7497	0.6648	0.3587	0.5372	0.7779	0.5222
MINIMDA	0.7503	0.4907	0.6788	0.7733	0.4105	0.5696	0.8102	0.6135
SMCLMDA	**0.8307**	**0.7026**	**0.8897**	**0.7992**	**0.6610**	**0.7852**	**0.8770**	**0.8010**

to concentrate on the most representative features and further enhance its predictive performance.

Key Points

- We organize the HMDD v4.0 and miR2Disease databases. The experimental data in this article includes causal/non-causal, up/down-regulation, and traditional association predictions, which basically cover all the relationships between miRNAs and diseases.
- We construct statistically significant meta-paths. These meta-paths are distinct from previous biological meta-paths in that statistical meta-paths can capture complex graph patterns and relationships without relying on biological the priori knowledge.
- We utilize the local embeddings generated by the Node2Vec algorithm as the initial features of GCN nodes. Additionally, the embedding representations learned from the meta-path view further strengthen those obtained from the similarity view through the contrastive learning method.
- To reflect the actual situation more realistically, we apply SMCLMDA to balanced and unbalanced test sets for validation. The results indicate that SMCLMDA outperforms other advanced methods. Additionally, the case study further verifies the robustness of SMCLMDA in predicting disease- related candidate miRNAs.

Competing interests

No competing interest is declared.

Acknowledgments

The authors thank the anonymous reviewers for their valuable suggestions.

Funding

This research received partial support from the National Natural Science Foundation of China (Grants 62162015 and 61762026), the Guangxi Natural Science Foundation (Grant 2023GXNSFAA026054), and the Innovation Project of Graduate Education at GUET (Grant 2024YCXB12).

Data Availability

https://github.com/yangfengzhuguet/SMCLMDA

References

1. DP Bartel. Micrornas: genomics, biogenesis, mechanism, and function. *Cell*, 166:281, 2014.
2. Eric A Miska. How micrornas control cell division, differentiation and death. *Current opinion in genetics & development*, 15(5):563–568, 2005.
3. Malwina Suszynska, Magdalena Machowska, Eliza Fraszczyk, Maciej Michalczyk, Anna Philips, Paulina Galka-Marciniak, and Piotr Kozlowski. Cmc: Cancer mirna census–a list of cancer-related mirna genes. *Nucleic Acids Research*, 52(4):1628–1644, 2024.
4. Chunmei Cui, Bitao Zhong, Rui Fan, and Qinghua Cui. Hmdd v4. 0: a database for experimentally supported human microrna-disease associations. *Nucleic Acids Research*, 52(D1):D1327–D1332, 2024.
5. Qinghua Jiang, Yadong Wang, Yangyang Hao, Liran Juan, Mingxiang Teng, Xinjun Zhang, Meimei Li, Guohua Wang, and Yunlong Liu. mir2disease: a manually curated database for microrna deregulation in human disease. *Nucleic acids research*, 37(suppl_1):D98–D104, 2009.
6. Boya Xie, Qin Ding, Hongjin Han, and Di Wu. mircancer: a microrna–cancer association database constructed by text mining on literature. *Bioinformatics*, 29(5):638–644, 2013.
7. Zhen Liu, Alhousseynou Sall, and Decheng Yang. Microrna: an emerging therapeutic target and intervention tool. *International journal of molecular sciences*, 9(6):978–999, 2008.
8. Zhengwei Li, Jiashu Li, Ru Nie, Zhu-Hong You, and Wenzheng Bao. A graph auto-encoder model for mirna-disease associations prediction. *Briefings in bioinformatics*, 22(4):bbaa240, 2021.
9. Xing Chen, Di Xie, Qi Zhao, and Zhu-Hong You. Micrornas and complex diseases: from experimental results to computational models. *Briefings in bioinformatics*, 20(2):515–539, 2019.
10. Xiangxiang Zeng, Xuan Zhang, and Quan Zou. Integrative approaches for predicting microrna function and prioritizing disease-related microrna using biological interaction networks. *Briefings in bioinformatics*, 17(2):193–203, 2016.
11. Yanyi Chu, Xuhong Wang, Qiuying Dai, Yanjing Wang, Qiankun Wang, Shaoliang Peng, Xiaoyong Wei, Jingfei Qiu, Dennis Russell Salahub, Yi Xiong, et al. Mda-gcnftg: identifying mirna-disease associations based on graph convolutional networks via graph sampling through the feature and topology graph. *Briefings in Bioinformatics*, 22(6):bbab165, 2021.
12. Cheng Yan, Jianxin Wang, Peng Ni, Wei Lan, Fang-Xiang Wu, and Yi Pan. Dnrlmf-mda: predicting microrna-disease associations based on similarities of micrornas and diseases. *IEEE/ACM transactions on computational biology and*

bioinformatics, 16(1):233–243, 2017.

13. Yu Han, Qiong Zhou, Leibo Liu, Jianwei Li, and Yuan Zhou. Dni-mdcap: improvement of causal mirna-disease association prediction based on deep network imputation. *BMC bioinformatics*, 25(1):22, 2024.
14. ShuDong Wang, YunYin Li, YuanYuan Zhang, ShanChen Pang, SiBo Qiao, Yu Zhang, and FuYu Wang. Generative adversarial matrix completion network based on multi-source data fusion for mirna–disease associations prediction. *Briefings in Bioinformatics*, 24(5):bbad270, 2023.
15. Liang Yu, Yujia Zheng, and Lin Gao. Mirna–disease association prediction based on meta-paths. *Briefings in Bioinformatics*, 23(2):bbab571, 2022.
16. Chunyu Wang, Kai Sun, Juexin Wang, and Maozu Guo. Data fusion-based algorithm for predicting mirna–disease associations. *Computational Biology and Chemistry*, 88:107357, 2020.
17. Dayun Liu, Yibiao Huang, Wenjuan Nie, Jiaxuan Zhang, and Lei Deng. Smalf: mirna-disease associations prediction based on stacked autoencoder and xgboost. *BMC bioinformatics*, 22(1):219, 2021.
18. Qiguo Dai, Zhaowei Wang, Ziqiang Liu, Xiaodong Duan, Jinmiao Song, and Maozu Guo. Predicting mirna-disease associations using an ensemble learning framework with resampling method. *Briefings in Bioinformatics*, 23(1):bbab543, 2022.
19. Chun-Chun Wang, Tian-Hao Li, Li Huang, and Xing Chen. Prediction of potential mirna–disease associations based on stacked autoencoder. *Briefings in bioinformatics*, 23(2):bbac021, 2022.
20. Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.
21. Wei Peng, Zhichen He, Wei Dai, and Wei Lan. Mhclmda: multihypergraph contrastive learning for mirna–disease association prediction. *Briefings in Bioinformatics*, 25(1):bbad524, 2024.
22. Qiao Ning, Yaomiao Zhao, Jun Gao, Chen Chen, Xiang Li, Tingting Li, and Minghao Yin. Amhmda: attention aware multi-view similarity networks and hypergraph learning for mirna–disease associations identification. *Briefings in Bioinformatics*, 24(2):bbad094, 2023.
23. Zhengzheng Lou, Zhaoxu Cheng, Hui Li, Zhixia Teng, Yang Liu, and Zhen Tian. Predicting mirna–disease associations via learning multimodal networks and fusing mixed neighborhood information. *Briefings in Bioinformatics*, 23(5):bbac159, 2022.
24. Xinru Tang, Jiawei Luo, Cong Shen, and Zihan Lai. Multi-view multichannel attention graph convolutional network for mirna–disease association prediction. *Briefings in Bioinformatics*, 22(6):bbab174, 2021.
25. Qiang He, Wei Qiao, Hui Fang, and Yang Bao. Improving the identification of mirna–disease associations with multi-task learning on gene–disease networks. *Briefings in Bioinformatics*, 24(4):bbad203, 2023.
26. Hsi-Yuan Huang, Yang-Chi-Dung Lin, Shidong Cui, Yixian Huang, Yun Tang, Jiatong Xu, Jiayang Bao, Yulin Li, Jia Wen, Huali Zuo, et al. mirtarbase update 2022: an informative resource for experimentally validated mirna–target interactions. *Nucleic acids research*, 50(D1):D222–D230, 2022.
27. Lynn M Schriml, Elvira Mitraka, James Munro, Becky Tauber, Mike Schor, Lance Nickle, Victor Felix, Linda Jeng, Cynthia Bearer, Richard Lichenstein, et al. Human disease ontology 2018 update: classification, content and workflow expansion. *Nucleic acids research*, 47(D1):D955–D962, 2019.
28. Carolyn E Lipscomb. Medical subject headings (mesh). *Bulletin of the Medical Library Association*, 88(3):265, 2000.
29. Guangzhan Zhang, Menglu Li, Huan Deng, Xinran Xu, Xuan Liu, and Wen Zhang. Sgnnmd: signed graph neural network for predicting deregulation types of mirna-disease associations. *Briefings in Bioinformatics*, 23(1):bbab464, 2022.
30. Wei Liu, Ting Tang, Xu Lu, Xiangzheng Fu, Yu Yang, and Li Peng. Mpclcda: predicting circrna–disease associations by using automatically selected meta-path and contrastive learning. *Briefings in Bioinformatics*, 24(4):bbad227, 2023.
31. Zhen Tian, Yue Yu, Haichuan Fang, Weixin Xie, and Maozu Guo. Predicting microbe–drug associations with structure-enhanced contrastive learning and self-paced negative sampling strategy. *Briefings in Bioinformatics*, 24(2):bbac634, 2023.
32. Ziming Liu, Yixuan Wang, Sachin Vaidya, Fabian Ruehle, James Halverson, Marin Soljačić, Thomas Y Hou, and Max Tegmark. Kan: Kolmogorov-arnold networks. *arXiv preprint arXiv:2404.19756*, 2024.
33. Zhen Yang, Liangcai Wu, Anqiang Wang, Wei Tang, Yi Zhao, Haitao Zhao, and Andrew E Teschendorff. dbdemc 2.0: updated database of differentially expressed mirnas in human cancers. *Nucleic acids research*, 45(D1):D812–D818, 2017.