


Article

Using Sequence Similarity Based on CKSNP Features and a Graph Neural Network Model to Identify miRNA–Disease Associations

Mingxin Li ^{1,†}, Yu Fan ^{1,†}, Yiting Zhang ^{2,3} and Zhibin Lv ^{1,*} ¹ College of Biomedical Engineering, Sichuan University, Chengdu 610065, China² College of Biology, Southwest Jiaotong University, Chengdu 611756, China³ College of Biology, Georgia State University, Atlanta, GA 30302-3965, USA

* Correspondence: lvzhibin@pku.edu.cn

† These authors contributed equally to this work.

Abstract: Among many machine learning models for analyzing the relationship between miRNAs and diseases, the prediction results are optimized by establishing different machine learning models, and less attention is paid to the feature information contained in the miRNA sequence itself. This study focused on the impact of the different feature information of miRNA sequences on the relationship between miRNA and disease. It was found that when the graph neural network used was the same and the miRNA features based on the K-spacer nucleic acid pair composition (CKSNAP) feature were adopted, a better graph neural network prediction model of miRNA–disease relationship could be built (AUC = 93.71%), which was 0.15% greater than the best model in the literature based on the same benchmark dataset. The optimized model was also used to predict miRNAs related to lung tumors, esophageal tumors, and kidney tumors, and 47, 47, and 37 of the top 50 miRNAs related to three diseases predicted separately by the model were consistent with descriptions in the wet experiment validation database (dbDEMOC).

Keywords: miRNA; miRNA sequence similarity; graph neural network; graph auto-encoder



Citation: Li, M.; Fan, Y.; Zhang, Y.; Lv, Z. Using Sequence Similarity Based on CKSNP Features and a Graph Neural Network Model to Identify miRNA–Disease Associations. *Genes* **2022**, *13*, 1759. <https://doi.org/10.3390/genes13101759>

Academic Editors: Quan Zou and Yijie Ding

Received: 4 August 2022

Accepted: 26 September 2022

Published: 28 September 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

According to molecular biology, the entire human genome can be divided into regulatory, coding, and noncoding genes. The coding genes are the main carriers of human genetic information, accounting for only 1.5% of the whole human genome [1–3]. Previous studies have regarded non-coding genes as transcriptional noise in the coding process. However, microarray experiments have found that non-coding genes can participate in cell activities by interacting with proteins and DNA, affecting gene activation and silencing; RNA splicing, modification, and editing; and protein translation, thus affecting various physiological processes. MicroRNA (miRNA) is a type of non-coding, single-stranded RNA encoded by endogenous genes. The most common role of miRNA is to directly regulate target genes by affecting post-transcriptional gene regulation of promoters. Abnormal expression of miRNA can cause many human diseases, and miRNA can be used as a drug target for disease treatment [4,5]. Therefore, the regulatory role of miRNA in disease expression is significant for a variety of complex human physiological processes and disease pathophysiology. However, traditional experimental methods are often limited. To this end, it is of great significance to speed up the verification process, reduce bias in biological experiments, and establish a method to predict the possible associations quickly and effectively between miRNAs and disease [6].

Common biological experimental methods from the 1990s to the beginning of the 21st century include reverse transcription-polymerase chain reaction (RT-PCR) [7], Northern blotting [8], and microarray profiling [9]. Although these traditional methods can accurately

detect correlations between miRNA and disease, they have their limitations. In recent years, a more classical class of algorithm has been developed based on the assumption of similarity measures [10]. This kind of algorithm is based on the assumption that miRNAs with similar functions are also related to similar disease phenotypes and can be sequenced accordingly [11]. Jiang et al. [12] constructed an miRNA association network using probability of interaction by target accessibility (PITA), and they proposed a hypergeometric distribution prediction method based on the human miRNA network. Liu et al. [13] proposed an miRNA–disease association prediction method by random walk on a heterogeneous network constructed by integrating multiple data sources. Zeng et al. [14] applied a link prediction algorithm named the structural perturbation method (SPM) on the miRNA–disease bilayer network to predict potential miRNA–disease associations. Zhang et al. [15] proposed a method using the correlation spectrum and the interaction network between miRNA and target genes to calculate the miRNA–disease association score by using fast linear neighborhood similarity-based network link inference. Mørk et al. [16] proposed a model to calculate the similarity between miRNA and disease based on the distance between target genes and disease genes in the PPI network. Another common method for predicting the miRNA–disease association is based on machine-inferred code similarity (MISIM) proposed by Wang et al. [17] Chen et al. [18] proposed a model called random forest for miRNA–disease association (RFMDA), which integrates disease semantic similarity, disease Gaussian interaction profile kernel similarity, miRNA Gaussian interaction profile kernel similarity, and miRNA functional similarity based on MISIM to predict miRNA–disease association. Zeng et al. [19] developed a neural network model to predict miRNA–disease associations (NNMDA). NNMDA not only aggregated the neighbor information during the process, but also preserved the topology of the original network at the same time. Zhou et al. [20] used similar methods to obtain miRNA and disease similarity networks and used the gradient boosting decision tree (GBDT) algorithm to extract more representative features. Additionally, with the recent great progress of graph neural networks in processing graph data, prediction methods based on graph neural networks have made breakthroughs, and the association between miRNA and disease is more suited to using graph neural networks than other data structures.

A key step in the data preprocessing for all models is to calculate the similarity of miRNAs. Specifically, these methods can be divided into three types: similarity measure-based methods, MISIM-based methods, and miRNA sequence similarity-based methods. Similarity measure-based methods construct an miRNA association network using target gene associations. However, this depends on the association between miRNA and target genes, resulting in high false positive and false negative rates [12,15,16,21–28]. The MISIM-based methods construct an miRNA association network using a disease semantic similarity network [17]. It can be said that related miRNAs have related diseases, but not all miRNAs related to similar diseases are necessarily related [18,20,28–30]. The miRNA similarity networks of such methods have the disadvantage of being dependent on known disease similarity networks, and this error is more pronounced for miRNAs that are associated with fewer diseases. The sequence similarity of miRNA is also particularly important for similarity networks. Previous studies have proposed enriching the miRNA similarity networks using miRNA sequence information. Ji et al. [31] integrated various kinds of information to construct a heterogeneous network centered on miRNA and disease, and embedded K-mer sequence features of miRNA into this network. Ji et al. [32] proposed a method using disease semantic similarity and miRNA sequence similarity to construct an miRNA–disease association network; miRNA sequence similarity-based methods effectively quantify the miRNA similarity, solving the problem that miRNA sequences cannot be directly compared due to their different lengths. In conclusion, according to current research, there is still room for improvement in the accuracy and effectiveness of identifying and predicting potential associations between miRNA and disease. Therefore, we enhanced the effective associations between miRNA and disease by using miRNA sequence information, benefit-

ting from the advantage of graph neural networks in finding miRNA–disease associations to construct a feature extraction method based on miRNA sequence similarity information.

In this study, we propose a method to calculate the sequence characteristics and sequence similarity of miRNA using five distinct miRNA sequence characteristics. Then, by integrating disease semantic similarity, miRNA Gaussian interaction profile kernel similarity, and disease Gaussian interaction profile kernel similarity, we construct a new bipartite graph of miRNA and disease. Then, the auto-encoder of a graph neural network is used to predict miRNA–disease association. Moreover, we evaluated prediction performance using 5-fold cross-validation. The model, using sequence similarity based on the composition of k-spaced nucleic acid pair (CKSNAP) features and a graph neural network, achieved an average area under the curve (AUC) of $93.71 \pm 0.42\%$, with an accuracy of $83.69 \pm 1.42\%$, precision of $77.73 \pm 2.29\%$, recall of $94.62 \pm 0.97\%$, and F1-score of $85.31 \pm 1.00\%$. To further verify the performance of this model, case studies on lung, esophageal, and kidney neoplasms were conducted. Respectively, the results showed that 47, 47, and 37 of the top 50 predicted miRNAs for these neoplasms can be confirmed by the database of Differentially Expressed MiRNAs in human Cancers (dbDEMC) [33]. In conclusion, it can be inferred that our model is effective and accurate in identifying potential associations between miRNA and disease.

2. Materials and Methods

2.1. Human miRNA–Disease Associations

In this study, we adopted the Human microRNA Disease Database (HMDD; v3.2) as the benchmark dataset and directly downloaded the experimentally verified miRNA–disease associations from <https://www.cuilab.cn/hmdd> (accessed on 1 September 2021) [34]. There are 16,427 high-quality miRNA–disease associations recorded in the HMDD database, including 877 diseases and 901 miRNAs. We adopted the adjacency matrix A to quantify associations between these miRNAs and diseases. If a disease is associated with an miRNA, the value of the element at the corresponding position of matrix A is set to 1, and otherwise to 0.

2.2. miRNA Sequence Similarity

In this study, the attribute features of miRNAs were represented by sequence similarity information. We downloaded miRNA sequence information from <https://mirbase.org/ftp.shtml> and utilized the K-mer ($k = 3$), Moran, Geary, NMBroto, and CKSNAP ($k = 5$) methods to obtain the sequence features of miRNAs.

K-mer: We set up a sliding window with a size of 3 and a sliding distance of 1 to obtain occurrence frequencies for all 3-monomere units. Then, each miRNA sequence was converted into a 64-dimensional vector based on the 64 3-monomer combinations. On this basis, we used cosine similarity, euclidean distance, and the Pearson correlation coefficient to calculate miRNA sequence similarity, defined as follows:

$$Kmer_cos_sim(i, j) = \frac{\sum_{j=1}^n \sum_{i=1}^n \frac{N(i)}{N} \times \frac{N(j)}{N}}{\sqrt{\sum_{i=1}^n \left(\frac{N(i)}{N}\right)^2} \times \sqrt{\sum_{j=1}^n \left(\frac{N(j)}{N}\right)^2}} \quad (1)$$

$$Kmer_euc_dist(i, j) = \sqrt{\sum_{i=1}^n \left(\frac{N(i)}{N} - \frac{N(j)}{N}\right)^2} \quad (2)$$

$$Kmer_pearson(i, j) = \frac{N \sum \frac{N(i)}{N} \times \frac{N(j)}{N} - \sum \frac{N(i)}{N} \sum \frac{N(j)}{N}}{\sqrt{N \sum \left(\frac{N(i)}{N}\right)^2 - \left(\sum \frac{N(i)}{N}\right)^2} \sqrt{N \sum \left(\frac{N(j)}{N}\right)^2 - \left(\sum \frac{N(j)}{N}\right)^2}} \quad (3)$$

where i represents a nucleotide combination with a length of three, such as AAA, AAC, and AAG; $N(i)$ is the number of nucleotide combinations and N is the length of a nucleotide sequence.

Moran: Moran describes miRNA by physicochemical parameters of nucleotides. The physical and chemical properties of miRNA include Rise (RNA), Roll(RNA), Shift(RNA), Slide(RNA), Tilt(RNA), Twist(RNA), Entropy(RNA), Adenine content, Purine(AG) content, Hydrophilicity(RNA), Enthalpy(RNA)1, GC content, Entropy(RNA)1, Hydrophilicity(RNA)1, Free energy(RNA), Keto(GT) content, Free energy(RNA)1, Enthalpy(RNA), Stacking energy(RNA), Guanine content, Cytosine content, and Thymine content. On this basis, we calculated the miRNA sequence similarity, defined as follows:

$$Moran_cos_sim(i, j) = \frac{\sum_{j=1}^n \sum_{i=1}^n M_i \times M_j}{\sqrt{\sum_{i=1}^n (M_i)^2} \times \sqrt{\sum_{j=1}^n (M_j)^2}} \tag{4}$$

$$Moran_euc_dist(i, j) = \sqrt{\sum_{i=1}^n \sum_{j=1}^n (M_i - M_j)^2} \tag{5}$$

$$Moran_pearson(i, j) = \frac{N \sum M_i \times M_j - \sum M_i \sum M_j}{\sqrt{N \sum M_i^2 - (\sum M_i)^2} \sqrt{N \sum M_j^2 - (\sum M_j)^2}} \tag{6}$$

where the Moran feature of miRNA are defined as follows:

$$M(d) = \frac{\frac{1}{N-d} \sum_{i=1}^{N-d} (M_i - \overline{M'}) (M_{i+d} - \overline{M'})}{\frac{1}{N} \sum_{i=1}^N (M_i - \overline{M'})^2}, d = 1, 2, 3, \dots, nlag \tag{7}$$

$$M = \frac{\sum_{r=1}^{len} \frac{M_r - \overline{M}}{\sqrt{\frac{1}{len} \sum_{r=1}^{len} (M_r - M)^2}}}{len} \tag{8}$$

where d represents the lag value of autocorrelation, and $nlag$ represents the maximum value of lag. In this study, the lag value $d = 3$ was selected; M_i is the nucleotide properties at position i . $\overline{M'}$ is calculated as follows:

$$\overline{M'} = \frac{\sum_{i=1}^N M_i}{N} \tag{9}$$

Geary: Geary uses the attribute information of nucleotides to describe the sequence characteristics of miRNA. We calculate the miRNA sequence similarity, defined as follows:

$$Geary_cos_sim(i, j) = \frac{\sum_{j=1}^n \sum_{i=1}^n G(d)_i \times G(d)_j}{\sqrt{\sum_{i=1}^n (G(d)_i)^2} \times \sqrt{\sum_{j=1}^n (G(d)_j)^2}} \tag{10}$$

$$Geary_euc_dist(i, j) = \sqrt{\sum_{i=1}^n \sum_{j=1}^n (G(d)_i - G(d)_j)^2} \tag{11}$$

$$Geary_pearson(i, j) = \frac{N \sum G(d)_i \times G(d)_j - \sum G(d)_i \sum G(d)_j}{\sqrt{N \sum G(d)_i^2 - (\sum G(d)_i)^2} \sqrt{N \sum G(d)_j^2 - (\sum G(d)_j)^2}} \tag{12}$$

where the Geary feature of miRNA are defined as follows:

$$G(d) = \frac{\frac{1}{2(N-d)} \sum_{i=1}^{N-d} (M_i - M_{i+d})^2}{\frac{1}{N-1} \sum_{i=1}^N (M_i - \bar{M}')^2}, d = 1, 2, \dots, nlag \quad (13)$$

The meaning of physical and chemical indicators involving nucleotides, such as d , M , M_i , and $nlag$, is the same as that in the Moran autocorrelation descriptor.

NMBroto: NMBroto is a normalized Moreau-Broto autocorrelation descriptor. The miRNA sequence is calculated similarity, defined as follows:

$$NMBroto_cos_sim(i, j) = \frac{\sum_{j=1}^n \sum_{i=1}^n NMB(d)_i \times NMB(d)_j}{\sqrt{\sum_{i=1}^n (NMB(d)_i)^2} \times \sqrt{\sum_{j=1}^n (NMB(d)_j)^2}} \quad (14)$$

$$NMBroto_euc_dist(i, j) = \sqrt{\sum_{j=1}^n (NMB(d)_i - NMB(d)_j)^2} \quad (15)$$

$$NMBroto_pearson(i, j) = \frac{N \sum NMB(d)_i \times NMB(d)_j - \sum NMB(d)_i \sum NMB(d)_j}{\sqrt{N \sum NMB(d)_i^2 - (\sum NMB(d)_i)^2} \sqrt{N \sum NMB(d)_j^2 - (\sum NMB(d)_j)^2}} \quad (16)$$

where the NMBroto feature of miRNA is defined as follows:

$$NMB(d) = \frac{\sum_{i=1}^{N-d} M_i \times M_{i+d}}{N-d}, d = 1, 2, \dots, nlag \quad (17)$$

The meaning of physical and chemical indicators involving nucleotides, such as d , M , M_i , and $nlag$, is the same as that in the Moran autocorrelation descriptor.

CKSNAP: CKSNAP uses k-spaced nucleic acid pairs to describe the frequency of separating the current nucleic acid pair from any k nucleic acids. When $k = 0$, there are 16 pairs of nucleic acid pairs with 0 spacing ('AA', 'AC', 'AG', 'AT', 'CA', 'CC', 'CG', 'CT', 'GA', 'GC', 'GG', 'GT', 'TA', 'TC', 'TG', 'TT'). The miRNA sequence was calculated similarity, defined as follows:

$$CKSNAP_cos_sim(i, j) = \frac{\sum_{j=1}^n \sum_{i=1}^n CK_i \times CK_j}{\sqrt{\sum_{i=1}^n (CK_i)^2} \times \sqrt{\sum_{j=1}^n (CK_j)^2}} \quad (18)$$

$$CKSNAP_euc_dist(i, j) = \sqrt{\sum_{j=1}^n (CK_i - CK_j)^2} \quad (19)$$

$$CKSNAP_pearson(i, j) = \frac{N \sum CK_i \times CK_j - \sum CK_i \sum CK_j}{\sqrt{N \sum CK_i^2 - (\sum CK_i)^2} \sqrt{N \sum CK_j^2 - (\sum CK_j)^2}} \quad (20)$$

where the CKSNAP features of miRNA are defined as follows:

$$CK = \left(\frac{N_{AA}}{N_{total}}, \frac{N_{AC}}{N_{total}}, \frac{N_{AG}}{N_{total}}, \dots, \frac{N_{TT}}{N_{total}} \right)_{16} \quad (21)$$

Using five miRNA sequence features and three similarity calculation methods, 15 miRNA sequence similarity matrices could be obtained. In the following research, all sequence similarity matrices are called MSSM.

2.3. Disease Semantic Similarity

Disease semantic similarity can be calculated based on the medical subject heading (MeSH) descriptors, which are available at <https://www.ncbi.nlm.nih.gov> (accessed on 1 January 2021). The relationship between diseases can be represented as a directed acyclic graph (DAG) network according to disciplines or affiliations, where the nodes represent the MeSH descriptors of diseases, and the directed edges point from parent nodes to child nodes [35]. We adopted $DAG_{d_i}(d_k)$ to describe semantic contribution of disease d_k to disease d_i . On this basis, the semantic contribution is defined as follows:

$$DAG_{d_i}(d_k) = \begin{cases} 1, & \text{if } d_k = d_i \\ \max\{\Delta * DAG_{d_i}(d'_k) \mid d'_k \in \text{children of } d_k\}, & \text{if } d_k \neq d_i \end{cases} \quad (22)$$

where Δ is the semantic contribution attenuation factor; this was set to 0.5 according to a previous study [30]. The semantic contribution value will decrease with distance.

According to the semantic contribution value of disease nodes, the semantic value of disease d_i was calculated as follows:

$$DV(d_i) = \sum_{d_k \in N(d_i)} DAG_{d_i}(d_k) \quad (23)$$

The semantic similarity between disease d_i and disease d_k could be calculated by the nodes shared by the two disease DAG networks. It was defined as follows:

$$SSM(d_i, d_j) = \frac{\sum_{d_t \in N(d_i) \cap N(d_j)} (DAG_{d_i}(d_t) + DAG_{d_j}(d_t))}{DV(d_i) + DV(d_j)} \quad (24)$$

where the element $DSSM(d_i, d_j)$ represents the disease semantic similarity between d_i and d_j .

2.4. Gaussian Interaction Profile Kernel Similarity for miRNAs and Diseases

In this study, the binary vector $IP(m_i)$ to denote the interaction profiles of miRNA m_i was defined by calculating the correlation between m_i and m_j . The miRNA Gaussian interaction profile kernel similarity matrix (MGSM) could be calculated as follows:

$$MGSM(m_i, m_j) = \exp\left(-\sigma_m \|IP(m_i) - IP(m_j)\|^2\right) \quad (25)$$

where σ_m was applied for controlling the bandwidth of the kernel, and $\|IP(m_i) - IP(m_j)\|^2$ was applied for calculating the euclidean distance of two eigenvectors. The Gaussian kernel bandwidth control parameter was calculated as follows:

$$\sigma_m = \sigma'_m / \left(\frac{1}{nm} \sum_{i=1}^{nm} \|IP(m_i)\|^2 \right) \quad (26)$$

where nm represents the number of all miRNAs; σ'_m was set to 1 according to a previous study [30]. Similarly, the Gaussian interaction profile kernel similarity for diseases $DGSM(d_i, d_j)$ between disease d_i and d_j could be calculated as follows:

$$DGSM(d_i, d_j) = \exp\left(-\sigma_d \|IP(d_i) - IP(d_j)\|^2\right) \quad (27)$$

$$\sigma_d = \sigma'_d / \left(\frac{1}{nd} \sum_{i=1}^{nd} \|IP(d_i)\|^2 \right) \quad (28)$$

where nd represents the number of all diseases, σ'_d was set to 1 according to a previous study [30].

2.5. Integrated Similarity for miRNAs and Diseases

There is a large number of sparse values in the miRNA sequence similarity and disease semantic similarity matrices. The final integrated similarity for miRNAs and diseases was obtained from miRNA sequence similarity, disease semantic similarity, Gaussian interaction profile kernel similarity for miRNAs, and Gaussian interaction profile kernel similarity for diseases. It was defined as follows:

$$MSim(m_i, m_j) = \begin{cases} MSSM(m_i, m_j), & \text{if } m_i \text{ and } m_j \text{ have sequence similarity} \\ MGSM(m_i, m_j), & \text{otherwise} \end{cases} \tag{29}$$

$$DSim(d_i, d_j) = \begin{cases} DSSM(d_i, d_j), & \text{if } d_i \text{ and } d_j \text{ have semantic similarity} \\ DGSM(d_i, d_j), & \text{otherwise} \end{cases} \tag{30}$$

2.6. Graph Auto-Encoder

In the present study, the prediction of the potential miRNA–disease associations was mainly used to generate the low-dimensional embedding of node information through the encoder based on a graph neural network, so as to realize the identification of the correlation between miRNA and diseases by a bilinear decoder. Our model can be described in four steps, as shown in Figure 1.

First, we constructed a bipartite graph including 877 disease nodes and 901 miRNA nodes. For integrated similarity for miRNA, the similarity between miRNA m_i and miRNA m_1, m_2, \dots, m_{901} could be expressed as follows:

$$F_m(i) = (u_1, u_2, u_3, \dots, u_{901}) \tag{31}$$

where $u_1, u_2, u_3, \dots, u_{901}$ is the integrated similarity between miRNA $m(i)$ and miRNA m_1, m_2, \dots, m_{901} . Similarly, the integrated similarity between disease d_i and disease d_1, d_2, \dots, d_{901} could be expressed as follows:

$$F_d(i) = (v_1, v_2, v_3, \dots, v_{877}) \tag{32}$$

The vectors in F_m (miRNA) and F_d (disease) were embedded into miRNA nodes and disease nodes in the miRNA–disease bipartite graph. In this study, 16,427 known association pairs verified by experiments were regarded as positive samples of input. In order to balance the positive and negative samples, 16,427 negative samples were randomly selected from the unknown associations in the subsequent experiments. In order to project miRNA and disease feature vectors to the same vector space, we designed a linear transformation matrix. The projection process of miRNA nodes can be described as follows:

$$H_m = W_{\varnothing_m} \cdot F_m \tag{33}$$

where F_m is the original characteristic matrix of miRNA, W_{\varnothing_m} is to realize the linear transformation matrix of miRNA projection process, and H_m is the feature matrix of miRNA projected into the new feature space. Similarly, the projection process of disease nodes can be described as follows:

$$H_d = W_{\varnothing_d} \cdot F_d \tag{34}$$

where F_d, W_{\varnothing_m} , and H_d are as described above.

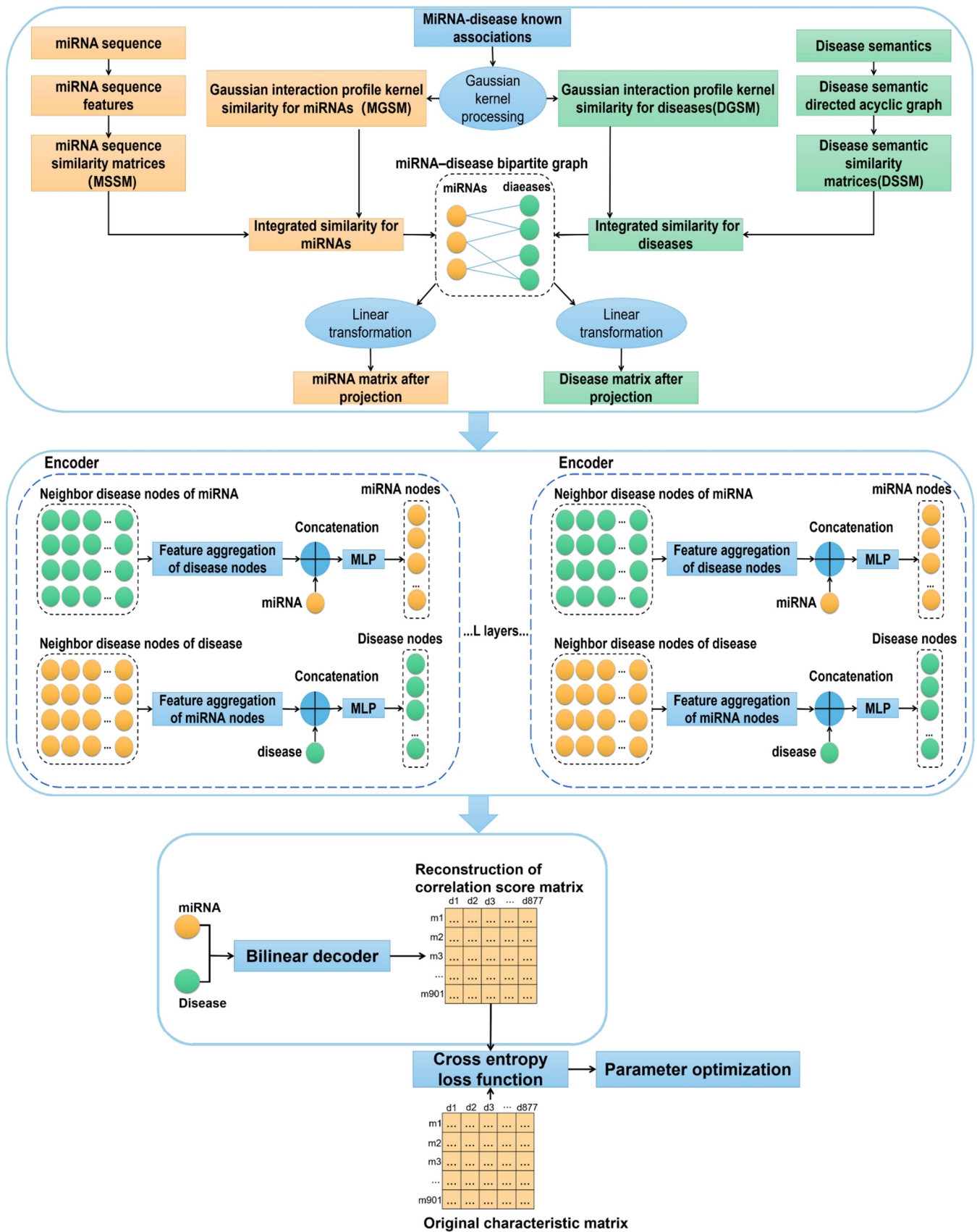


Figure 1. Technical route of miRNA disease correlation prediction model.

Second, we used the aggregator function of the auto-encoder to integrate the feature representation of the node and its neighbor nodes, update the node with a multi-layer perceptron, and embed the aggregated features into the original features of the node. The aggregator function $sum(\cdot)$ was used to realize the feature aggregation of its neighbor nodes, as shown in Equations (35) and (36), where $sum(H_d(1), H_d(2), \dots)$ represents the aggregation process of the disease nodes d_j , which are the direct neighbors of the miRNA nodes m_i ; D_{m_i} is the degree value of node m_i . In order to prevent numerical instability, $H_m^s(i)$ was used for normalization. It is the aggregation feature of the current miRNA node m_i .

$$H_m^s(i) = \frac{1}{D_{m_i}} sum(H_d(1), H_d(2), \dots) \tag{35}$$

$$D_{m_i} = |\{d_j | \exists e_{ij} \in E \text{ or } e_{ji} \in E\}| \tag{36}$$

After the aggregation features of all miRNA nodes were obtained through the aggregator function, the features of all nodes were embedded and connected through the multi-layer perceptron superimposed by the L-layer, and the final embedding of nodes was generated. We used a Leaky Rectified Linear Unit (LeakyReLU) function as the activation function of the multilayer perceptron to avoid the phenomenon of neuron “death” when the input was negative. It is defined as follows:

$$H'_m(i) = LeakyReLU(f(H_m(i) \oplus H_m^s(i))) \tag{37}$$

where $H_m(i)$ is the original feature of node m_i , $H_m^s(i)$ is the aggregation feature of node m_i , $H'_m(i)$ is the updated node feature, and $f(\cdot)$ is the single-layer multi-layer perceptron function. Similarly, we calculated the aggregation feature of the current disease node d_j , the degree value of node d_j , and the updated node feature of disease as in Equations (38)–(40).

$$H_d^s(j) = \frac{1}{D_d(j)} sum(H_m(1), H_m(2), \dots) \tag{38}$$

$$D_d(j) = |\{m_i | \exists e_{ij} \in E \text{ or } e_{ji} \in E\}| \tag{39}$$

$$H'_d(j) = LeakyReLU(f(H_d(j) \oplus H_d^s(j))) \tag{40}$$

A multilayer overlay network could enhance the feature embedding of neighbor nodes and preserve the topology of graph data, so as to enhance the expression ability of features.

Third, considering that the number of unknown associations between miRNA and disease is much larger than the known associations, we adopted a sigmoid function as the activation function of the decoder to predict the association score between miRNA and disease. It is defined as follows:

$$\hat{A}(i, j) = sigmoid\left(H_d^L(j) * Q\left(H_m^L(i)\right)^T\right) \tag{41}$$

where Q is the E-dimensional trainable parameter matrix, $H_d^L(j)$ is the feature embedding of L-layer disease node d_j , $H_m^L(i)$ is the feature embedding of L-layer miRNA node m_i , and $\hat{A}(i, j)$ is the reconstructed association score matrix.

Finally, we used the deviation between the prediction score matrix and the original characteristic matrix and used the cross-entropy loss function and back propagation algorithm to optimize the model to obtain the best model parameters. It is defined as follows:

$$LOSS = - \sum_{i,j \in y \cup y^-} (A(i, j) * \log \hat{A}(i, j) + (1 - A(i, j)) * \log(1 - \hat{A}(i, j))) \tag{42}$$

where $A(i, j)$ is the original characteristic matrix, and $\hat{A}(i, j)$ is the reconstructed prediction correlation score matrix. Due to the small proportion of known associations in the sample data, cost-sensitive learning was used to improve the weight of positive sample loss, so as

to improve the prediction accuracy. Furthermore, y and y^- are the set of positive samples and the set of negative samples in the incidence matrix, respectively.

2.7. Model Evaluation

Five-fold cross validation was selected to evaluate the performance of the model. We divided the miRNA–disease associations into five sets, set one as the test set and the other four as the training set, and received five prediction results. Accordingly, the higher the score between miRNA and disease, the higher the possibility of a potential association between them. In the present study, we adopted four common evaluating indicators to evaluate the performance of the models: Accuracy (Acc), Precision (Prec), Recall, and F1-score. Meanwhile, we plotted the receiver operating characteristic curves to intuitively display the performance of our model and utilized the AUC to comprehensively evaluate model performance.

3. Results

3.1. Performance Evaluation of Graph Neural Network Prediction Model Based on Single Features

Based on five sequence features, we used three different similarity calculation methods to obtain 15 different miRNA sequence similarity matrices and constructed 15 prediction models. The performance comparison of the prediction models based on different characteristics and similarity calculation methods is shown in Figure 2. The overall prediction performance of the model for calculating sequence feature similarity based on the Pearson correlation coefficient was better than the other two similarity calculation methods. This type of model obtained the optimal values in AUC, ACC, precision, and F1-score, and had obvious advantages over the other two similarity algorithms. Comparing the performance of the model based on five features, we found that the performance of the model based on K-mer and CKSNAP was better than the other models. Considering that AUC could more comprehensively evaluate model performance, the model based on CKSNAP sequence features and the Pearson similarity calculation method obtained the highest AUC value among the 15 models. In addition, this model was also better than other models in ACC, precision, and F1 scores.

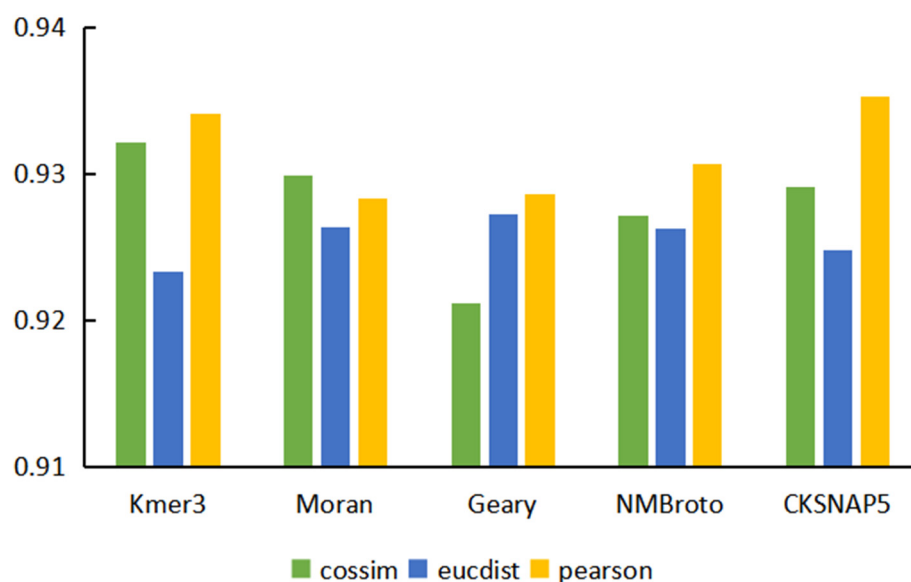


Figure 2. AUC of models based on different features and different similarity calculation methods.

3.2. Performance Evaluation of Graph Neural Network Prediction Model Based on Combined Features

Based on the comparison and analysis of model performance based on single features, we proposed prediction models based on combined features, hoping to enhance

the expression ability of nodes by embedding multiple features on a single node. Here, we combined the five features in pairs to obtain 10 combined features and used different similarity calculation methods to build 30 different prediction models based on these combined features. In order to compare the performance of the two types of models more intuitively, we classified these two types of models and calculated their average scores on each evaluation indicator and drew the prediction performance based on single-feature and double-feature models, as shown in Figure 3. We noticed that the evaluation indicators of models based on combined features were lower than those of the models based on single features. We speculate that this is because the dimensions of the combined features were too high. Embedding more feature information into a single node also introduced too much noise and redundant information, which led to model instability in the subsequent experimental process, resulting in a decline in prediction performance.

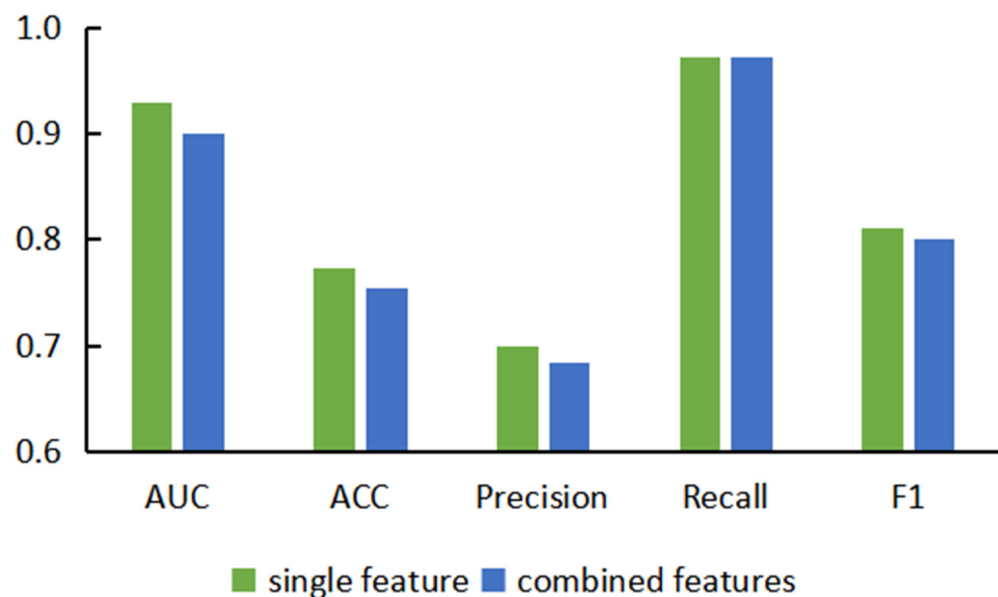


Figure 3. Performance comparison diagram of combined feature-based models and single feature-based models.

In addition, we also found that the score of the models based on combined features depended on the score of the models based on single features. Here, we selected six groups of models for display, as shown in Figure 4. It can be seen that the models based on single features were better than the models based on combined features in AUC, Acc, Prec, and F1-score. In Figure 4, the gray columns are mostly located below the other two columns. The recall rate of models based on combined features in individual groups is higher. The reason for this is that the combined features enrich the information range, which leads to the improvement of recall rate. Combined features could enrich the information contained in a single node, which would make the coverage of information and the construction of relationship structure in the whole network more comprehensive, resulting in the improvement of Recall. However, due to the high dimensions of the feature vector and too much noise information on a single node, other evaluation indicators would decline.

Based on this, the subsequent experiments focused on performance optimization based on single-feature models. In a preliminary study, we set project dimension to $E = 256$ and set encoder layers to $L = 2$.

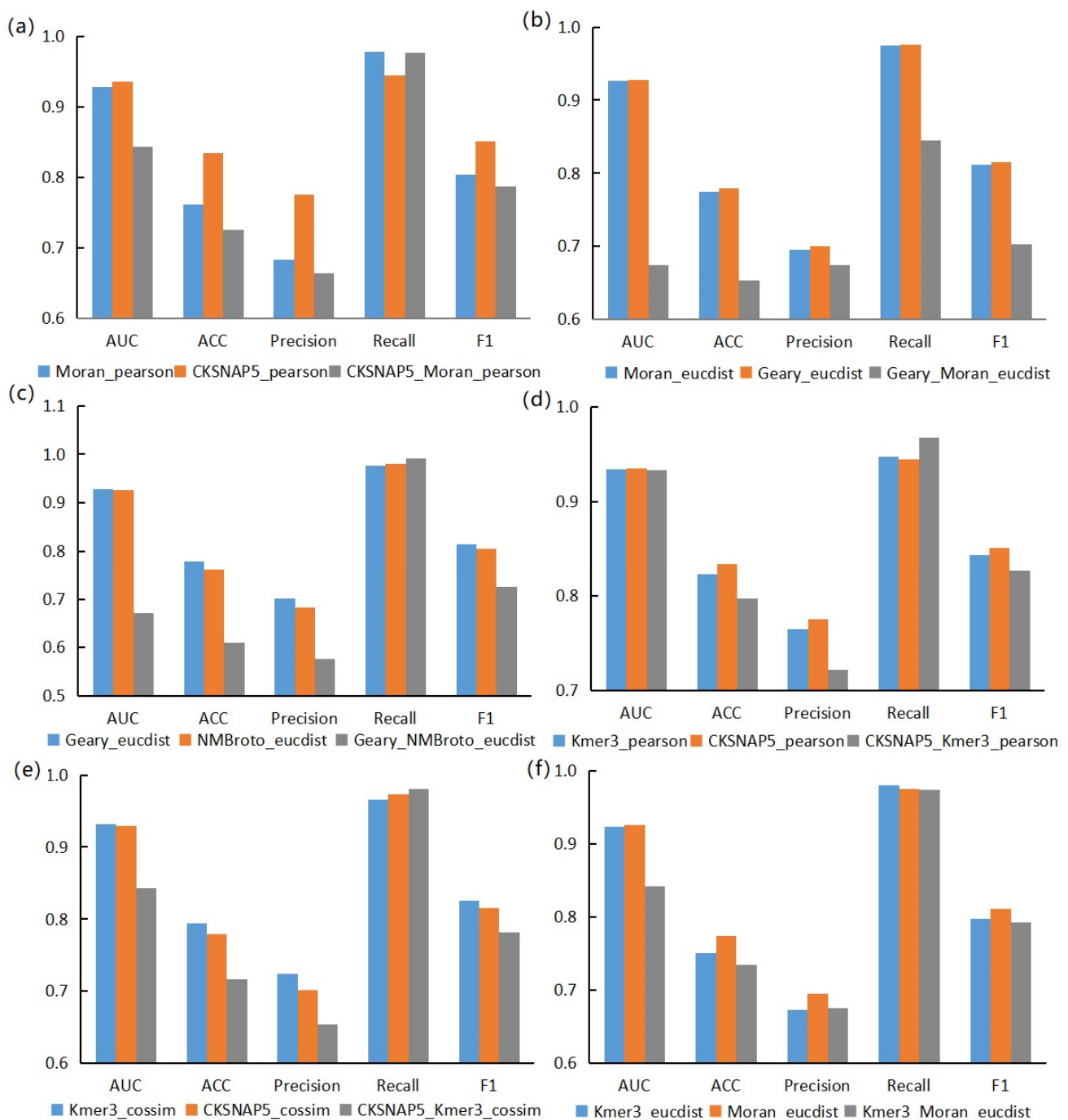


Figure 4. Example diagram of dependence of dual feature models on single feature models. (a) is a comparison between the Pearson correlation coefficient model based on Moran feature (blue), the Pearson correlation coefficient model based on CKSNAP feature (dark orange) and the Pearson correlation coefficient model based on Moran+CKSNAP combined feature (gray). (b) is a comparison between the Euclidean distance model based on Moran feature (blue), the Euclidean distance model based on Geary feature (dark orange) and the Euclidean distance model based on Moran+Geary combined feature (gray). (c) is a comparison between the Euclidean distance model based on Geary feature (blue), the Euclidean distance model based on NMBroto feature (dark orange) and the Euclidean distance model based on Geary+NMBroto combined feature (gray). (d) is a comparison between the Pearson correlation coefficient model based on Kmer feature (blue), the Pearson

correlation coefficient model based on CKSNAP feature (dark orange) and the Pearson correlation coefficient model based on Kmer+CKSNAP combined feature (gray). (e) is a comparison between the cosine similarity model based on Kmer feature (blue) and the cosine similarity model based on CKSNAP feature (yellow) and the cosine similarity model based on Kmer+CKSNAP combined feature (gray). (f) is a comparison between the Euclidean distance model based on Kmer feature (blue), the Euclidean distance model based on Moran feature (dark orange) and the Euclidean distance model based on Kmer+Moran combined feature (gray).

3.3. Effects of Projection Dimension and Encoder Layers on Model Performance

Here, we used the model based on CKSNAP features and the Pearson similarity calculation method to study the influence of projection dimension and encoder layers on prediction performance, changing the number of projection dimensions and encoder layers to obtain the scores of different models on five evaluation indicators. Different projection dimensions will affect the expression ability of nodes. Changing the number of encoder layers can adjust the learning degree of neural networks to node feature information. As can be seen from Figure 5a, with an increase in projection dimensions, the recall rate of the model fluctuated greatly, and the change in F1 score was not obvious. The accuracy score increased significantly with the increase of projection dimensions but dropped sharply when the projection dimension, E , was greater than 256. According to Figure 5b, with an increase in encoder layers, the evaluation indexes of the model showed a downward trend as a whole; especially after $L > 6$, the indexes of the model declined sharply.

3.4. Performance Evaluation and Comparative Analysis of Related Models

Here, we marked the top five models in each evaluation indicator and screened out the models with two or more markers to obtain the results shown in Figure 6. Considering that Acc was greatly affected by the proportion of positive samples in the sample set, Acc was not taken as the primary evaluation indicator. Precision and recall provide a single view of the performance, whereas F1 score provides a more comprehensive view of the performance and therefore was used for model evaluation. Therefore, considering AUC and F1 score, we noted that the model with project dimensions $E = 64$ and encoder layer $L = 6$ had better performance in all aspects. The receiver operating characteristic curve for the 5-fold cross validation experiment is shown in Figure 7. At the same time, the ROC curve of each fold can be seen in Figure S1.

In order to further evaluate the performance of this model, we compared it with six related models (PBMDA [36], LLCMDA [37], EDTMDA [38], GBDTLR [20], MCLPMDA [39], GAEMDA [40]) for comprehensive evaluation. Considering that different studies used different evaluation indicators, only the AUC value that could comprehensively evaluate the performance of the model was selected for comparative analysis. Our study selected the optimal AUC recorded in each paper for comparison, as shown in Table 1. Among the seven models, our model obtained the highest AUC value, which was 0.15% higher than the AUC value of the model with the second highest, GAEMDA.

Table 1. Comparative analysis of our model and related models.

Model	AUC (%)
PBMDA	91.72
LLCMDA	91.90
EDTMDA	91.92
GBDTLR	92.74
MCLPMDA	93.20
GAEMDA	93.56
Our Model	93.71

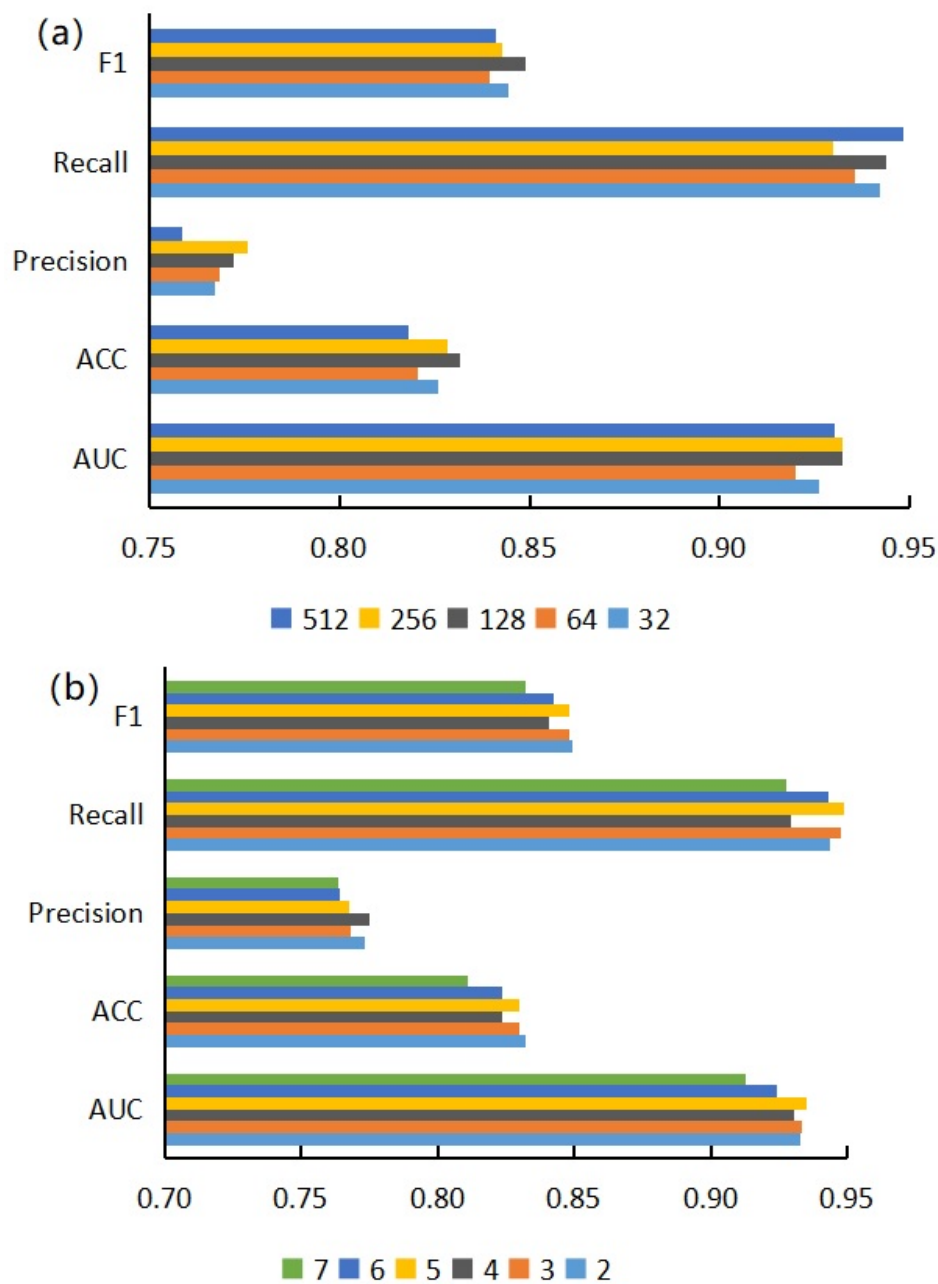


Figure 5. (a) Effects of different projection dimensions on model performance; (b) influence of different encoder layers on model performance.

3.5. Case Studies

In recent years, more and more research has shown that the mutation or abnormal expression of miRNA causes many human diseases [41,42]. In order to further evaluate the performance of our prediction algorithm, three neoplasm diseases were selected for independent case studies—lung neoplasm, esophageal neoplasm, and kidney neoplasm—as it done in the reported method using the same dataset. We deleted the specific diseases of the case study from the training samples to remove bias from the experiments. We used the remaining miRNAs and diseases to construct test samples and ranked them according to the prediction scores. We compared the top 50 prediction results with the dbDEMC databases to obtain the prediction results.

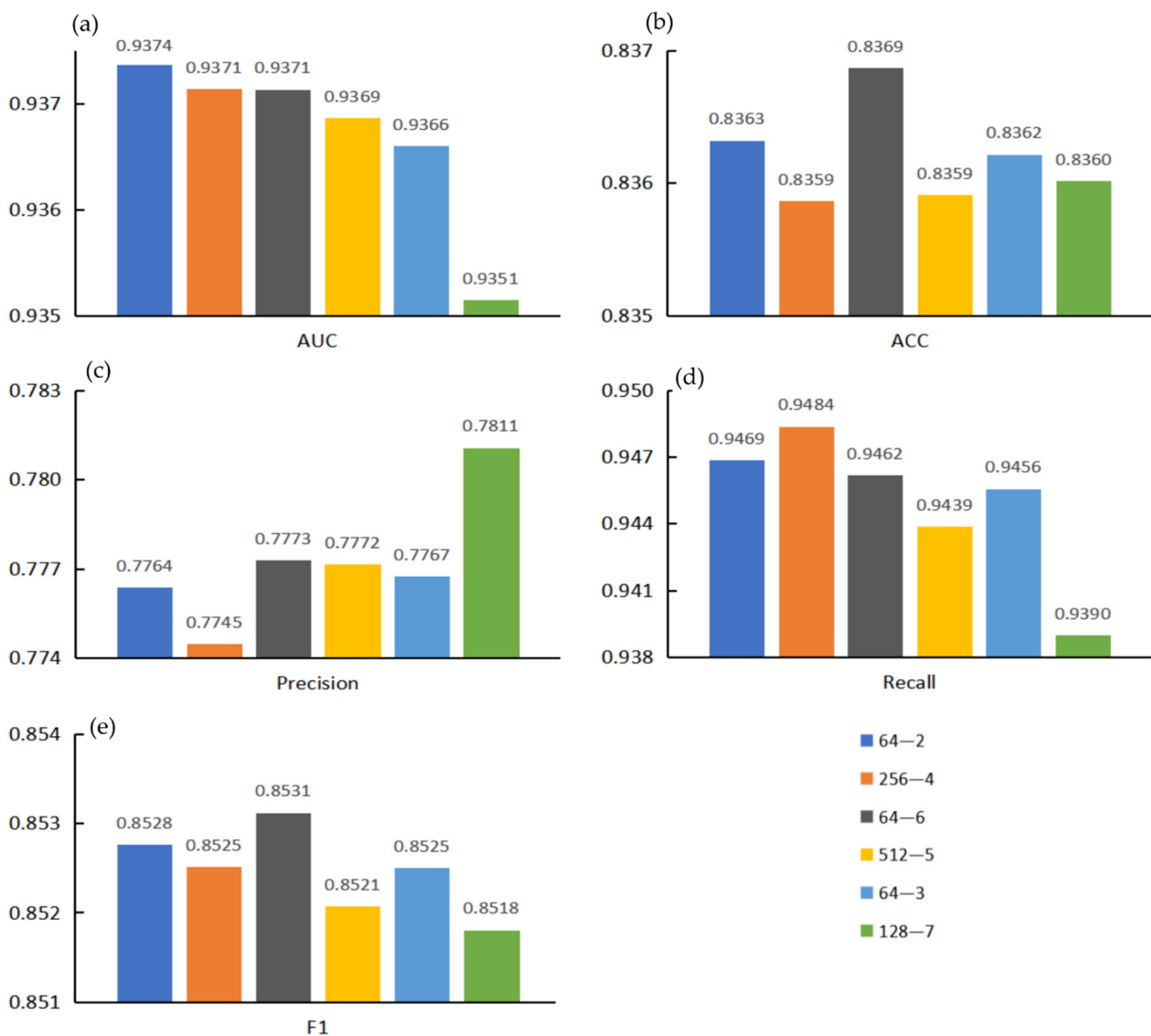


Figure 6. The top five evaluation indicators were marked and summarized to obtain two or more models. (a) AUC value comparison; (b) ACC value comparison; (c) Precision value comparison; (d) Recall value comparison; (e) F1 value comparison.

Lung neoplasm is a malignant neoplasm disease occurring in lung parenchyma and stroma [43]. Lung cancer is one of the diseases with the fastest growth rate of morbidity and mortality. It has become the most common cause of death in malignant tumors; the prediction results of our model for miRNA related to lung neoplasm are shown in Table 2. It can be seen that 47 of the top 50 miRNAs could be verified in the dbDEMC database. Esophageal neoplasm is a malignant neoplasm disease that occurs in esophageal epithelial tissue [44–46]. Approximately 300,000 people die of esophageal cancer every year in the world. China is one of the high incidence areas of esophageal cancer in the world; the prediction results of our model for miRNA related to esophageal neoplasm are shown in Table 3. It can be seen that 47 of the top 50 miRNAs could be verified in the dbDEMC database. Kidney neoplasm is a common neoplasm disease in the urinary system. The pathological structure of kidney neoplasm is complex, and the cause of disease is variable [47]. Finding a new entry point for diagnosis is of great significance for the timely targeting of patients. The prediction results of our model are shown in Table 4. It can be seen that 37 of the top 50 miRNAs could be verified in the dbDEMC database.

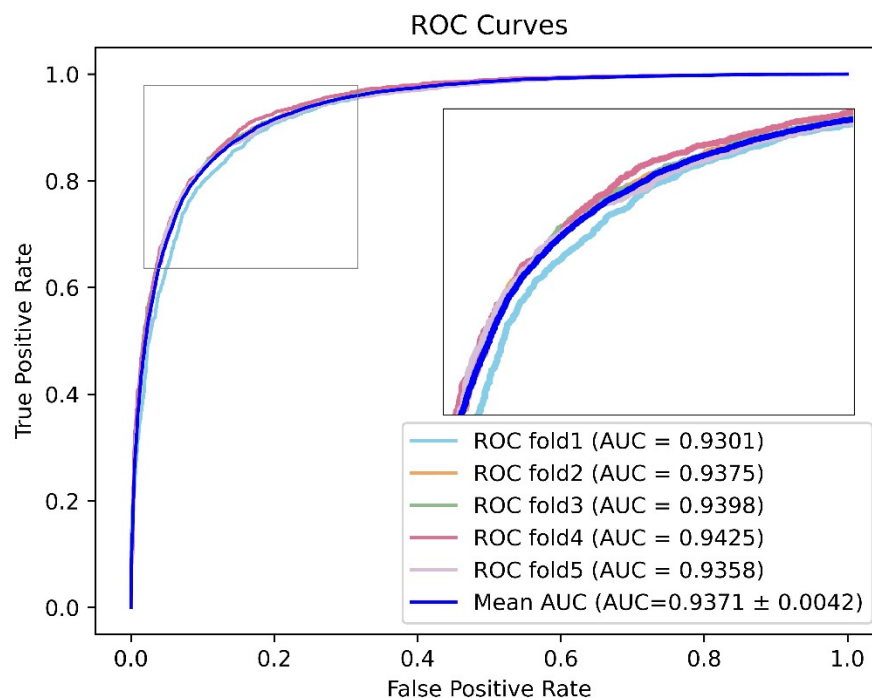


Figure 7. The model with project dimension E = 64 and encoder layers L = 6 had better performance.

Table 2. Lung neoplasm associated miRNA Top 50 predicted by hmdd v3.2.

miRNA	dbDEMC	miRNA	dbDEMC
hsa-mir-586	Confirmed	hsa-mir-329-5p	Confirmed
hsa-mir-208b-5p	Confirmed	hsa-mir-1264	Confirmed
hsa-mir-376b-5p	Confirmed	hsa-mir-618	Confirmed
hsa-mir-3613-5p	Confirmed	hsa-mir-599	Confirmed
hsa-mir-4775	Confirmed	hsa-mir-517c-3p	Unconfirmed
hsa-mir-544a	Confirmed	hsa-mir-384	Confirmed
hsa-mir-450a-5p	Confirmed	hsa-mir-581	Confirmed
hsa-mir-376c-5p	Confirmed	hsa-mir-578	Confirmed
hsa-mir-376a-5p	Confirmed	hsa-mir-19b-2-5p	Confirmed
hsa-mir-190a-5p	Confirmed	hsa-mir-552-5p	Confirmed
hsa-mir-875-5p	Confirmed	hsa-mir-5590-5p	Confirmed
hsa-mir-3682-5p	Confirmed	hsa-mir-450a-1-3p	Confirmed
hsa-mir-302f	Confirmed	hsa-mir-454-5p	Confirmed
hsa-mir-5586-5p	Confirmed	hsa-mir-942-5p	Confirmed
hsa-mir-450b-5p	Confirmed	hsa-mir-548l	Confirmed
hsa-mir-576-5p	Confirmed	hsa-mir-548k	Confirmed
hsa-mir-4295	Confirmed	hsa-mir-1185-5p	Confirmed
hsa-mir-1282	Confirmed	hsa-mir-548am-5p	Confirmed
hsa-mir-5009-5p	Confirmed	hsa-mir-613	Confirmed
hsa-mir-655-5p	Confirmed	hsa-mir-1248	Confirmed
hsa-mir-16-2-3p	Confirmed	hsa-mir-544b	Confirmed
hsa-mir-548d-5p	Confirmed	hsa-mir-3913-5p	Confirmed
hsa-mir-1179	Confirmed	hsa-mir-548c-5p	Confirmed
hsa-mir-876-5p	Confirmed	hsa-mir-570-5p	Unconfirmed
hsa-mir-1206	Unconfirmed	hsa-mir-651-5p	Confirmed

Table 3. Esophageal neoplasm associated miRNA Top 50 predicted by hmdd v3.2.

miRNA	dbDEMOC	miRNA	dbDEMOC
hsa-mir-1179	Confirmed	hsa-mir-450b-5p	Confirmed
hsa-mir-1206	Confirmed	hsa-mir-4775	Confirmed
hsa-mir-1264	Confirmed	hsa-mir-493-5p	Confirmed
hsa-mir-1282	Confirmed	hsa-mir-495-5p	Confirmed
hsa-mir-135a-5p	Confirmed	hsa-mir-5009-5p	Confirmed
hsa-mir-136-5p	Confirmed	hsa-mir-517c-3p	Confirmed
hsa-mir-16-2-3p	Confirmed	hsa-mir-544a	Confirmed
hsa-mir-190a-5p	Confirmed	hsa-mir-545-5p	Confirmed
hsa-mir-196a-5p	Confirmed	hsa-mir-548d-5p	Confirmed
hsa-mir-199b-5p	Confirmed	hsa-mir-552-5p	Unconfirmed
hsa-mir-19b-2-5p	Confirmed	hsa-mir-5586-5p	Confirmed
hsa-mir-202-5p	Confirmed	hsa-mir-5590-5p	Confirmed
hsa-mir-208b-5p	Confirmed	hsa-mir-576-5p	Confirmed
hsa-mir-29a-5p	Confirmed	hsa-mir-578	Confirmed
hsa-mir-329-5p	Unconfirmed	hsa-mir-581	Confirmed
hsa-mir-3613-5p	Confirmed	hsa-mir-586	Confirmed
hsa-mir-3682-5p	Confirmed	hsa-mir-599	Confirmed
hsa-mir-376a-2-5p	Confirmed	hsa-mir-618	Confirmed
hsa-mir-376a-5p	Confirmed	hsa-mir-655-5p	Confirmed
hsa-mir-376c-5p	Confirmed	hsa-mir-7-5p	Confirmed
hsa-mir-384	Confirmed	hsa-mir-875-5p	Confirmed
hsa-mir-4295	Confirmed	hsa-mir-876-5p	Confirmed
hsa-mir-4423-5p	Confirmed	hsa-mir-95-5p	Confirmed
hsa-mir-450a-1-3p	Unconfirmed	hsa-mir-9-5p	Confirmed
hsa-mir-450a-5p	Confirmed	hsa-mir-29b-1-5p	Confirmed

Table 4. Kidney neoplasm associated miRNA Top 50 predicted by hmdd v3.2.

miRNA	dbDEMOC	miRNA	dbDEMOC
hsa-mir-105-5p	Confirmed	hsa-mir-449a	Confirmed
hsa-mir-1179	Confirmed	hsa-mir-449c-5p	Confirmed
hsa-mir-1204	Confirmed	hsa-mir-4775	Confirmed
hsa-mir-1244	Confirmed	hsa-mir-4795-5p	Unconfirmed
hsa-mir-1264	Confirmed	hsa-mir-517c-3p	Confirmed
hsa-mir-1267	Confirmed	hsa-mir-5193	Unconfirmed
hsa-mir-1282	Confirmed	hsa-mir-520h	Unconfirmed
hsa-mir-1284	Confirmed	hsa-mir-543	Confirmed
hsa-mir-1322	Confirmed	hsa-mir-548c-5p	Unconfirmed
hsa-mir-135b-5p	Confirmed	hsa-mir-5692b	Unconfirmed
hsa-mir-136-5p	Confirmed	hsa-mir-576-5p	Confirmed
hsa-mir-147b-5p	Unconfirmed	hsa-mir-577	Confirmed
hsa-mir-149-5p	Confirmed	hsa-mir-586	Confirmed
hsa-mir-18b-5p	Confirmed	hsa-mir-606	Confirmed
hsa-mir-202-5p	Confirmed	hsa-mir-616-5p	Confirmed
hsa-mir-212-5p	Confirmed	hsa-mir-626	Unconfirmed
hsa-mir-23c	Confirmed	hsa-mir-633	Confirmed
hsa-mir-3120-5p	Unconfirmed	hsa-mir-644a	Unconfirmed
hsa-mir-3149	Confirmed	hsa-mir-645	Confirmed
hsa-mir-32-5p	Unconfirmed	hsa-mir-764	Unconfirmed
hsa-mir-340-5p	Confirmed	hsa-mir-889-5p	Unconfirmed
hsa-mir-3662	Confirmed	hsa-mir-934	Confirmed
hsa-mir-3682-5p	Unconfirmed	hsa-mir-942-5p	Confirmed
hsa-mir-4295	Confirmed	hsa-mir-943	Confirmed
hsa-mir-4443	Confirmed	hsa-mir-944	Confirmed

4. Conclusions

In the present study, we used a variety of miRNA sequence features to better retain the sequence similarity information between miRNAs and used a disease–miRNA bipartite graph to mine for potential deeper association information between miRNAs and diseases. Under a 5-fold cross validation, the experimental results showed that the prediction performance of the prediction algorithm based on combined features was not as good as that based on single features, and there was a dependency between the prediction score based on combined features and the prediction model based on corresponding single features. We used this model for case study verification, and the prediction results of three specific tumors also achieved a good hit rate. Therefore, our research is helpful for researchers to quickly and effectively study the relationship between miRNAs and diseases and plays a guiding role in research. It can save time and the cost of wet experiments to find disease-relevant miRNAs. This method can be used to predict the miRNA associated with a disease, and then perform wet experiment verification. Alternatively, there are results from wet experiments, and the method could be used to provide a confidence to refer to in experimental results. However, the analysis and discussion described in this paper is only a small part of the research on the correlation between miRNAs and disease, and there are more questions to be explored. For example, it could be useful to embed more biological information in the structural association between disease and miRNA. For example, when calculating the sequence similarity of miRNAs, we can consider introducing the functional similarity of miRNAs, the MISIM network, and the correlation information between miRNAs and proteins. Further work can focus on finding methods that can obtain deep-seated network structure information without affecting the prediction performance of the model. Finally, considering that the regulatory mechanisms of miRNA in many complex diseases also play an important role in miRNA–disease associations, the analysis of model performance combined with physiological influencing factors is expected to improve future experiments.

Supplementary Materials: The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/genes13101759/s1>, Figure S1. Receiver operating characteristic curve for each fold experiment. The source code can be downloaded from http://public.aibiochem.net/DNA_RNA/Genes_Human-miRNA-disease-Associations/code.zip (accessed on 1 January 2022).

Author Contributions: Conceptualization, Z.L.; methodology, Z.L.; software, Z.L. and Y.F.; writing—original draft preparation, Y.F. and M.L.; writing—review and editing, Y.F., Y.Z., and M.L.; supervision, Z.L.; funding acquisition, Z.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by National Natural Science Foundation of China (nos. 62001090) and Fundamental Research Funds for the Central Universities of Sichuan University (nos. YJ2021104).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Crick, F.H.C.; Barnett, L.; Brenner, S.; Watts-Tobin, R.J. General Nature of the Genetic Code for Proteins. *Nature* **1961**, *192*, 1227–1232. [[CrossRef](#)] [[PubMed](#)]
2. Yanofsky, C. Establishing the Triplet Nature of the Genetic Code. *Cell* **2007**, *128*, 815–818. [[CrossRef](#)] [[PubMed](#)]
3. Bertone, P.; Stolc, V.; Royce, T.E.; Rozowsky, J.S.; Urban, A.E.; Zhu, X.; Rinn, J.L.; Tongprasit, W.; Samanta, M.; Weissman, S. Global identification of human transcribed sequences with genome tiling arrays. *Science* **2004**, *306*, 2242–2246. [[CrossRef](#)] [[PubMed](#)]
4. Mishra, R.; Bhattacharya, S.; Rawat, B.S.; Kumar, A.; Kumar, A.; Niraj, K.; Chande, A.; Gandhi, P.; Khetan, D.; Aggarwal, A. MicroRNA-30e-5p has an integrated role in the regulation of the innate immune response during virus infection and systemic lupus erythematosus. *iScience* **2020**, *23*, 101322. [[CrossRef](#)]

5. Tang, W.; Wan, S.; Yang, Z.; Teschendorff, A.E.; Zou, Q. Tumor Origin Detection with Tissue-Specific miRNA and DNA methylation Markers. *Bioinformatics* **2018**, *34*, 398–406. [[CrossRef](#)] [[PubMed](#)]
6. Wong, L.; You, Z.-H.; Guo, Z.-H.; Yi, H.-C.; Chen, Z.-H.; Cao, M.-Y. MIPDH: A Novel Computational Model for Predicting microRNA–mRNA Interactions by DeepWalk on a Heterogeneous Network. *ACS Omega* **2020**, *5*, 17022–17032. [[CrossRef](#)]
7. Freeman, W.; Walker, S.; Vrana, K. Quantitative RT-PCR: Pitfalls and potential. *BioTechniques* **1999**, *26*, 112–122; 124–125. [[CrossRef](#)]
8. Várallyay, É.; Burgyán, J.; Havelda, Z. MicroRNA detection by northern blotting using locked nucleic acid probes. *Nat. Protoc.* **2008**, *3*, 190–196. [[CrossRef](#)]
9. Baskerville, S.; Bartel, D.P. Microarray profiling of microRNAs reveals frequent coexpression with neighboring miRNAs and host genes. *Rna* **2005**, *11*, 241–247. [[CrossRef](#)]
10. 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. [[CrossRef](#)]
11. Chen, X.; Xie, D.; Zhao, Q.; You, Z.H. MicroRNAs and complex diseases: From experimental results to computational models. *Brief. Bioinform.* **2019**, *20*, 515–539. [[CrossRef](#)] [[PubMed](#)]
12. Jiang, Q.; Hao, Y.; Wang, G.; Juan, L.; Zhang, T.; Teng, M.; Liu, Y.; Wang, Y. Prioritization of disease microRNAs through a human phenome-microRNAome network. *BMC Syst. Biol.* **2010**, *4* (Suppl. S1), S2. [[CrossRef](#)] [[PubMed](#)]
13. Liu, Y.; Zeng, X.; He, Z.; Zou, Q. Inferring microRNA-disease associations by random walk on a heterogeneous network with multiple data sources. *IEEE ACM Trans. Comput. Biol. Bioinform.* **2017**, *14*, 905–915. [[CrossRef](#)] [[PubMed](#)]
14. Zeng, X.; Liu, L.; Lu, L.; Zou, Q. Prediction of potential disease-associated microRNAs using structural perturbation method. *Bioinformatics* **2018**, *34*, 2425–2432. [[CrossRef](#)]
15. Zhang, W.; Li, Z.; Guo, W.; Yang, W.; Huang, F. A fast linear neighborhood similarity-based network link inference method to predict microRNA-disease associations. *IEEE ACM Trans. Comput. Biol. Bioinform.* **2019**, *18*, 405–415. [[CrossRef](#)]
16. Mørk, S.; Pletscher-Frankild, S.; Pallega Caro, A.; Gorodkin, J.; Jensen, L.J. Protein-driven inference of miRNA–disease associations. *Bioinformatics* **2013**, *30*, 392–397. [[CrossRef](#)]
17. Wang, D.; Wang, J.; Lu, M.; Song, F.; Cui, Q. Inferring the human microRNA functional similarity and functional network based on microRNA-associated diseases. *Bioinformatics* **2010**, *26*, 1644–1650. [[CrossRef](#)]
18. Chen, X.; Wang, C.C.; Yin, J.; You, Z.H. Novel Human miRNA-Disease Association Inference Based on Random Forest. *Mol. Ther. Nucleic Acids* **2018**, *13*, 568–579. [[CrossRef](#)]
19. Zeng, X.; Wang, W.; Deng, G.; Bing, J.; Zou, Q. Prediction of potential disease-associated microRNAs by using neural network. *Mol. Ther. Nucleic Acids* **2019**, *16*, 566–575. [[CrossRef](#)]
20. Zhou, S.; Wang, S.; Wu, Q.; Azim, R.; Li, W. Predicting potential miRNA-disease associations by combining gradient boosting decision tree with logistic regression. *Comput. Biol. Chem.* **2020**, *85*, 107200. [[CrossRef](#)]
21. Ma, Y.; He, T.; Ge, L.; Zhang, C.; Jiang, X. MiRNA-disease interaction prediction based on kernel neighborhood similarity and multi-network bidirectional propagation. *BMC Med. Genom.* **2019**, *12*, 185. [[CrossRef](#)] [[PubMed](#)]
22. Ping, X.; Ke, H.; Guo, M.; Guo, Y.; Li, J.; Jian, D.; Yong, L.; Dai, Q.; Jin, L.; Teng, Z. Correction: Prediction of microRNAs Associated with Human Diseases Based on Weighted k Most Similar Neighbors. *PLoS ONE* **2013**, *8*, 3752034. [[CrossRef](#)]
23. Lu, X.; Gao, Y.; Zhu, Z.; Ding, L.; Wang, X.; Liu, F.; Li, J. A Constrained Probabilistic Matrix Decomposition Method for Predicting miRNA-disease Associations. *Curr. Bioinform.* **2021**, *16*, 524–533. [[CrossRef](#)]
24. Tian, L.; Wang, S.-L. Exploring miRNA Sponge Networks of Breast Cancer by Combining miRNA-disease-lncRNA and miRNA-target Networks. *Curr. Bioinform.* **2021**, *16*, 385–394. [[CrossRef](#)]
25. Zhang, J.; Sun, Q.; Liang, C. Prediction of lncRNA-disease Associations Based on Robust Multi-label Learning. *Curr. Bioinform.* **2021**, *16*, 1179–1189. [[CrossRef](#)]
26. Zhang, Y.; Duan, G.; Yan, C.; Yi, H.; Wu, F.-X.; Wang, J. MDAPLatform: A Component-based Platform for Constructing and Assessing miRNA-disease Association Prediction Methods. *Curr. Bioinform.* **2021**, *16*, 710–721. [[CrossRef](#)]
27. Zhu, Q.; Fan, Y.; Pan, X. Fusing Multiple Biological Networks to Effectively Predict miRNA-disease Associations. *Curr. Bioinform.* **2021**, *16*, 371–384. [[CrossRef](#)]
28. Jiang, L.; Zhu, J. Review of MiRNA-disease association prediction. *Curr. Protein Pept. Sci.* **2020**, *21*, 1044–1053. [[CrossRef](#)]
29. Chen, X.; Yan, C.C.; Zhang, X.; You, Z.-H.; Deng, L.; Liu, Y.; Zhang, Y.; Dai, Q. WBSMDA: Within and between score for MiRNA-disease association prediction. *Sci. Rep.* **2016**, *6*, 21106. [[CrossRef](#)]
30. Yao, D.; Zhan, X.; Kwok, C.-K. An improved random forest-based computational model for predicting novel miRNA-disease associations. *BMC Bioinform.* **2019**, *20*, 624. [[CrossRef](#)]
31. Ji, B.-Y.; You, Z.-H.; Cheng, L.; Zhou, J.-R.; Alghazzawi, D.; Li, L.-P. Predicting miRNA-disease association from heterogeneous information network with GraRep embedding model. *Sci. Rep.* **2020**, *10*, 6658. [[CrossRef](#)] [[PubMed](#)]
32. Ji, B.-Y.; You, Z.-H.; Wang, Y.; Li, Z.-W.; Wong, L. DANE-MDA: Predicting microRNA-disease associations via deep attributed network embedding. *iScience* **2021**, *24*, 102455. [[CrossRef](#)]
33. Yang, Z.; Ren, F.; Liu, C.; He, S.; Sun, G.; Gao, Q.; Yao, L.; Zhang, Y.; Miao, R.; Cao, Y.; et al. dbDEMC: A database of differentially expressed miRNAs in human cancers. *BMC Genom.* **2010**, *11*, S5. [[CrossRef](#)] [[PubMed](#)]
34. Huang, Z.; Shi, J.; Gao, Y.; Cui, C.; Zhang, S.; Li, J.; Zhou, Y.; Cui, Q. HMDD v3.0: A database for experimentally supported human microRNA–disease associations. *Nucleic Acids Res.* **2018**, *47*, D1013–D1017. [[CrossRef](#)] [[PubMed](#)]

35. Chen, X.; Clarence Yan, C.; Luo, C.; Ji, W.; Zhang, Y.; Dai, Q. Constructing lncRNA functional similarity network based on lncRNA-disease associations and disease semantic similarity. *Sci. Rep.* **2015**, *5*, 11338. [[CrossRef](#)] [[PubMed](#)]
36. You, Z.-H.; Huang, Z.-A.; Zhu, Z.; Yan, G.-Y.; Li, Z.-W.; Wen, Z.; Chen, X. PBMDA: A novel and effective path-based computational model for miRNA-disease association prediction. *PLoS Comput. Biol.* **2017**, *13*, 1005455. [[CrossRef](#)] [[PubMed](#)]
37. Qu, Y.; Zhang, H.; Lyu, C.; Liang, C. LLCMDA: A Novel Method for Predicting miRNA Gene and Disease Relationship Based on Locality-Constrained Linear Coding. *Front. Genet.* **2018**, *9*, 576. [[CrossRef](#)] [[PubMed](#)]
38. Chen, X.; Zhu, C.-C.; Yin, J. Ensemble of decision tree reveals potential miRNA-disease associations. *PLoS Comput. Biol.* **2019**, *15*, 1007209. [[CrossRef](#)]
39. Yu, S.P.; Liang, C.; Xiao, Q.; Li, G.H.; Ding, P.J.; Luo, J.W. MCLPMDA: A novel method for miRNA-disease association prediction based on matrix completion and label propagation. *J. Cell. Mol. Med.* **2019**, *23*, 1427–1438. [[CrossRef](#)]
40. Li, Z.; Li, J.; Nie, R.; You, Z.-H.; Bao, W. A graph auto-encoder model for miRNA-disease associations prediction. *Brief. Bioinform.* **2020**, *22*, bbaa240. [[CrossRef](#)]
41. Birks, D.K.; Barton, V.N.; Donson, A.M.; Handler, M.H.; Vibhakar, R.; Foreman, N.K. Survey of MicroRNA expression in pediatric brain tumors. *Pediatr. Blood Cancer* **2011**, *56*, 211–216. [[CrossRef](#)] [[PubMed](#)]
42. Alder, H.; Taccioli, C.; Chen, H.; Jiang, Y.; Smalley, K.J.; Fadda, P.; Ozer, H.G.; Huebner, K.; Farber, J.L.; Croce, C.M.; et al. Dysregulation of miR-31 and miR-21 induced by zinc deficiency promotes esophageal cancer. *Carcinogenesis* **2012**, *33*, 1736–1744. [[CrossRef](#)] [[PubMed](#)]
43. Torre, L.A.; Siegel, R.L.; Jemal, A. Lung Cancer Statistics. *Adv. Exp. Med. Biol.* **2016**, *893*, 1–19. [[CrossRef](#)]
44. Linehan, W.M. Genetic basis of kidney cancer: Role of genomics for the development of disease-based therapeutics. *Genome Res.* **2012**, *22*, 2089–2100. [[CrossRef](#)]
45. Senanayake, U.; Das, S.; Vesely, P.; Alzoughbi, W.; Fröhlich, L.F.; Chowdhury, P.; Leuschner, I.; Hoefler, G.; Guertl, B. miR-192, miR-194, miR-215, miR-200c and miR-141 are downregulated and their common target ACVR2B is strongly expressed in renal childhood neoplasms. *Carcinogenesis* **2012**, *33*, 1014–1021. [[CrossRef](#)] [[PubMed](#)]
46. Zaman, M.S.; Shahryari, V.; Deng, G.; Thamminana, S.; Saini, S.; Majid, S.; Chang, I.; Hirata, H.; Ueno, K.; Yamamura, S. Correction: Up-Regulation of MicroRNA-21 Correlates with Lower Kidney Cancer Survival. *PLoS ONE* **2012**, *7*, 31060. [[CrossRef](#)]
47. Kim, K.; Taylor, S.L.; Ganti, S.; Guo, L.; Weiss, R.H. Urine Metabolomic Analysis Identifies Potential Biomarkers and Pathogenic Pathways in Kidney Cancer. *Omics A J. Integr. Biol.* **2011**, *15*, 293–303. [[CrossRef](#)]