biomedicines

MDPI

Article

Heterogeneous Types of miRNA-Disease Associations Stratified by Multi-Layer Network Embedding and Prediction

Dong-Ling Yu [1,2], Zu-Guo Yu [1,2,*], Guo-Sheng Han [1,2], Jinyan Li [3,*] and Vo Anh [4]

[1] Key Laboratory of Intelligent Computing and Information Processing of Ministry of Education, Xiangtan University, Xiangtan 411105, China; 201931000105@smail.xtu.edu.cn (D.-L.Y.); hangs@xtu.edu.cn (G.-S.H.)
[2] Hunan Key Laboratory for Computation and Simulation in Science and Engineering, Xiangtan University, Xiangtan 411105, China
[3] Data Science Institute, University of Technology Sydney, Broadway, NSW 2007, Australia
[4] Faculty of Science, Engineering and Technology, Swinburne University of Technology, Hawthorn, VIC 3122 , Australia; vanh@swin.edu.au
* Correspondence: yuzuguo@aliyun.com (Z.-G.Y.); Jinyan.Li@uts.edu.au (J.L.)

Abstract: Abnormal miRNA functions are widely involved in many diseases recorded in the database of experimentally supported human miRNA-disease associations (HMDD). Some of the associations are complicated: There can be up to five heterogeneous association types of miRNA with the same disease, including genetics type, epigenetics type, circulating miRNAs type, miRNA tissue expression type and miRNA-target interaction type. When one type of association is known for an miRNA-disease pair, it is important to predict any other types of the association for a better understanding of the disease mechanism. It is even more important to reveal associations for currently unassociated miRNAs and diseases. Methods have been recently proposed to make predictions on the association types of miRNA-disease pairs through restricted Boltzman machines, label propagation theories and tensor completion algorithms. None of them has exploited the non-linear characteristics in the miRNA-disease association network to improve the performance. We propose to use attributed multi-layer heterogeneous network embedding to learn the latent representations of miRNAs and diseases from each association type and then to predict the existence of the association type for all the miRNA-disease pairs. The performance of our method is compared with two newest methods via 10-fold cross-validation on the database HMDD v3.2 to demonstrate the superior prediction achieved by our method under different settings. Moreover, our real predictions made beyond the HMDD database can be all validated by NCBI literatures, confirming that our method is capable of accurately predicting new associations of miRNAs with diseases and their association types as well.

Keywords: miRNA-disease; heterogeneous association types; attributed multi-layer heterogeneous network embedding; Node2vec

check for updates

Citation: Yu, D.-L.; Yu, Z.-G.; Han, G.-S.; Li, J.; Anh, V. Heterogeneous Types of miRNA-Disease Associations Stratified by Multi-Layer Network Embedding and Prediction. *Biomedicines* **2021**, *9*, 1152. https://doi.org/10.3390/biomedicines9091152

Academic Editosr: Santiago Garcia-Vallve and Francesca Lovat

Received: 29 April 2021
Accepted: 30 August 2021
Published: 3 September 2021

1. Introduction

MicroRNAs (miRNAs) are a class of non-coding single-stranded RNA molecules with a length of about 22 nucleotides encoded from endogenous genes. It has been found that the abnormally high or low expressions of miRNAs are closely related to disease progression and development, such as tumor progression [1,2]. Therefore, miRNAs can be used as biomarkers for disease diagnosis or drug targets in treatment design [3,4]. These associations between miRNAs and diseases are sometimes very complicated with up to five heterogeneous types induced by genetics, epigenetics, circulating miRNAs, miRNA tissue expression and miRNA-target interactions.

An early version of the database HMDD (v2.0) [5] stratifies miRNA-disease associations into four types (denoted as Type-1, Type-2, Type-3 and Type-4 here). Type-1 is defined as a special association induced by genetics. It is mainly confirmed by GWAS analysis, gene knockout and gene overexpression. For example, two SNPs of rs41275794 and rs12976445 in pri-miR-125a change the expression of mature miR-125a and are associated with recurrent abortion [6]; the deletion of miR-15a increases the risk of Lymphoma [7]; the proliferation, invasion and metastasis of HCT116 cells can be inhibited by overexpression of miR-34a in colon cancer [8]. Type-2 is a special association induced by epigenetics change. Mutations of epigenetic components are common in diseases [9]. It has been proved that miRNAs can serve on modulators of epigenetic through targeting key enzymes that are responsible for epigenetic responses, and the expression of miRNAs can be regulated by epigenetic mechanism [10]. The reciprocal interaction between epigenetic and miRNA regulation forms an epigenetic-miRNA feedback loop. Type-3 association is specified from circulation assays. Since circulating miRNAs are always remarkably stable in plasma and serum under harsh conditions, they can be used as novel biomarkers for diagnosis and prognosis [11,12]. Circulating miRNAs can also be selectively targeted for secretion in one cell and absorbed by a distant target cell to regulate gene expression [13]. Type-4 can be identified through miRNA-target interactions, including miRNA-mRNA interactions and miRNA-lncRNA interactions, as well as feedback loops between miRNAs and transcription factors. For instance, miR-106a targeting MCL1 to inhibit cisplatin resistance of A2780 in ovarian cancer cells [14] implies a Type-4 association between miR-106a and ovarian cancer; the feedback loops between miR-124 and TGF-β pathways play an important role in metastasis of non-small cell lung cancer [15]. Recently, the HMDD database was updated to version v3.2 [16] and added a newly specified classification of Type-5 on the evidence and data from miRNA tissue expression assays. MiRNAs are present in a variety of human tissues, and the differential expression of miRNA and the deregulation expression of miRNA in tissues have been found to be related to the disease. For example, it has been found that six miRNAs, miR-31, miR-34a, miR-181a, miR-181b, miR-193a-3p and miR-193b are all upregulated, but other the four miRNAs miR-221, miR-222, miR-484 and miR-502-3p are all downregulated in pancreatic cancer cells [17]. Moreover, Cui et al. [18] explored a significant positive correlation among body fluid miRNAs and tissue miRNAs. The miRNAs in tissues are highly correlated with miRNAs in male serum and female plasma.

Therefore, the association of miRNA-disease can not only be divided into multiple types according to the association evidences but there are also correlations among the various types. Furthermore, when an miRNA is associated with a disease, the association can be multiple types. For example, the reduced expression level of miR-9 can not only be the prognosis biomarker for Waldenstrom macroglobulinemia but it also causes Waldenstrom macroglobulinemia in humans. The association between miR-223 and inflammation can be proved by genetics, epigenetics or miRNA-target interactions. When one type of association is known for an miRNA-disease pair, it is important to predict any other types of the association for a better understanding of the disease mechanism and for developing accurate treatment for the diseases. It is even more important to reveal associations and association type for currently unassociated miRNAs and diseases. Since confirming associations and association types of miRNA-disease through biological experiments is very time-consuming and costly, it is very necessary to reveal miRNA-disease association and association type through computational methods.

Most of previous algorithms aimed for a binary prediction to observe whether an miRNA-disease pair has an association or not, but these binary prediction algorithms are unable to stratify the multiple heterogeneous types of the association between an miRNA and a disease. To refine the prediction extendable to handle the multiple types of an association, Chen et al. [19] proposed a model (RBMMMDA) based on restricted Boltzman machines (RBMs) where a RBM is constructed for every specific miRNA. All of the diseases are then served as nodes in the visible layer, and the associated state with a specific miRNA acts as the label of the disease node. The association probabilities between a disease and an miRNA are updated by a two-layer undirected graph. RBMMMDA can predict the links of miRNAs associated with diseases as well as multiple association types. However, the visible layers and the hidden layers are not cross connected in

RBM, missing the similarity information of the disease network and the miRNA network. Zhang et al. [20] proposed to calculate the Gaussian similarity and semantic similarity for diseases, the function similarity and the Gaussian similarity for miRNAs on each miRNA-disease association type. Then, a model (NLPMMDA) was constructed to uncover unobserved multiple types of miRNA-disease associations. For each association type, they adopted a label propagation algorithm on the integrated disease-disease similarity network and corresponding miRNA-miRNA similarity network. Good performance was achieved by NLPMMDA, whereas the correlation between multi-types of miRNA-disease associations was not taken into consideration. Recently, Huang et al. [21] considered multiple association types between miRNAs and diseases as a tensor and developed a novel tensor completion algorithm (named TDRC) to predict miRNA-disease association types. TDRC treats auxiliary information of miRNA similarities and disease similarities as relational constraints and incorporates it into the optimal objective function. As TDRC takes into account multiple biological information and the correlation between different association types, it performs better than previous models and other traditional tensor completion algorithms in exploiting potential miRNA-disease association types. However, the tensor completion algorithm can only capture linear relationships of disease-miRNA association network. We believe that it would be very effective to predict the miRNA-disease multi-type associations by using the non-linear properties among the disease-miRNA associations, the miRNA similarity network and the disease similarity network.

Recent development in graph embedding or network representation learning have opened the door for exploring non-linear properties in heterogeneous networks. The technique of attributed multi-layer heterogeneous network embedding (GATNE) [22] is a recently emerging graph embedding approach which is able to capture rich attributed information and exploit multiplex topological structures from different node types. In particular, it can project the structural information of the nodes and non-linear relationship of the network into low-dimensional continuous space while preserving inherent properties. Considering that the network of miRNA-disease is complicated, comprising not only multi-typed nodes and associations but also rich similarity information between the nodes of the same type, we are motivated to explore attributed multi-layer heterogeneous network embedding for predictions of multi-typed associations between miRNAs and diseases. We name our method mDLinker.

By our method, the node features of miRNAs and diseases are created through the Node2vec algorithm from their similarity networks. The topology information, the non-linear relationships of miRNA-disease associations and different association types are captured through GATNE from the attributed multi-layer miRNA-disease heterogeneous network. Finally, random forest is trained on the information-rich edges of the network and then applied to predict miRNA-disease association types within the HMDD database or beyond the database. Experimental results show that our method achieved excellent performance in both algorithm comparison and real case studies. Therefore, our algorithm has powerful ability to predict miRNA-disease association and association types. It also has great significance for understanding disease pathology at the molecular level.

2. Materials and Methods

2.1. Datasets

Data records from HMDD, MeSH and miRBase were downloaded for this study. HMDD (http://www.cuilab.cn/hmdd/, accessed on 11 October 2020) is a well-maintained database of miRNA-disease associations. HMDD v2.0 recorded 10,381 miRNA-disease associations between 383 diseases and 577 miRNAs, while the latest version of HMDD v3.2 covers 894 diseases, 1206 miRNAs and 35,548 miRNA-disease associations. The miRBase database (http://www.mirbase.org/, accessed on 15 October 2020) is one of the most important public databases of miRNAs, which provides published miRNA precursor sequences, their annotations, predicted gene targets and so on. The directed acyclic graph description of diseases is obtained from the Medical Subject Heading (MeSH) database in the National Library of

Medicine (http://www.nlm.nih.gov/, accessed on 14 October 2020). For our study, miRNAs and diseases with irregular or incorrect names were deleted. Due to the sparsity of miRNA-disease association types in the miRNA-disease matrix, similarly as [20], we mapped different miRNA precursors into the same mature miRNAs in HMDD v2.0. A summary of the data records in HMDD v2.0 and HMDD v3.2 is presented in Table A1. Figure 1 shows the proportions of associations containing 1, 2, 3, 4 or 5 types.

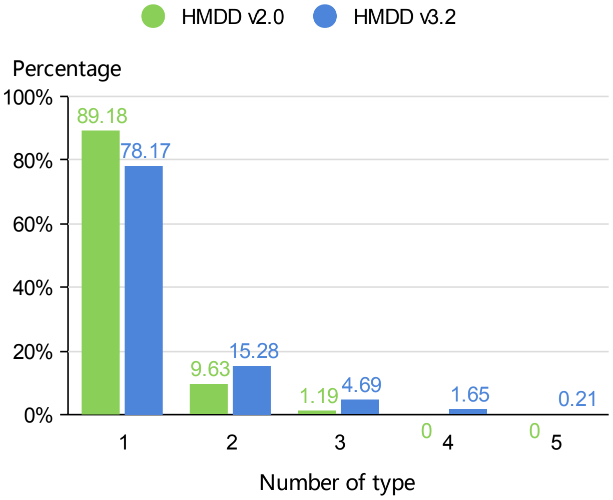


Figure 1. The percentages of miRNA-disease associations containing different numbers of types.

2.2. Similarity Calculation for Diseases and miRNAs

The calculation method for the semantic similarity of diseases was proposed by Wang et al. [23]. In the descriptor C of the MeSH database, the relationships among the diseases are represented as directed acyclic graphs (DAGs), which are usually directed from a general disease to a more specific disease. An example of the DAG structure of infectious mononucleosis is shown in Figure 2, where the directionality of the edges is used to find ancestor nodes of infectious mononucleosis, and more distant ancestors contributed less semantically to infectious mononucleosis. For disease P, the semantic values $DV(P)$ can be calculated as follows:

$$DV(P) = \sum_{t \in N_P} D_P(t) \tag{1}$$

$$\begin{cases} D_P(t) = max\, \alpha * D_P(t\prime)|t\prime \in children(t), \quad if \quad t \neq P \\ D_P(P) = 1 \end{cases} \tag{2}$$

where N_P is the node set including ancestor nodes of P and disease P. α is the ancestor contribution factor $\alpha = 0.5$ suggested by [23]. since the two diseases with more ancestor nodes are more similar, the semantic similarity between disease P and disease Q can be calculated as follows.

$$sim(P,Q) = \frac{\sum_{t \in N_P \cap N_Q}(D_P(t) + D_Q(t))}{DV(P) + DV(Q)} \tag{3}$$

Based on the assumption that similar miRNA precursor sequences will have similar functions, miRNA precursor sequences are effectively used to extract miRNA features or to calculate miRNA similarity in the prediction of miRNA-disease binary association. For example, Li et al. [24] and Ji et al. [25] adopted the 3-mer method to construct miRNAs feature; Che et al. [26] measured the miRNA-miRNA similarity by the Levenshtein distance of miRNA precursor sequences; Zheng et al. [27] calculated an miRNA–miRNA sequence similarity matrix by using a chaos game representation (CGR). In order to make effective use of biological structural information of miRNAs and avoid the recalculation of traditional miRNA function similarity in cross-validation, the miRNAs similarity calculation method proposed by Zheng et al. [27] is adopted in our study where miRNA precursor sequences are mapped to a Euclidean space by an iterative mapping function, and then the region distance of CGRs is calculated to measure the similarity between miRNA and miRNA.

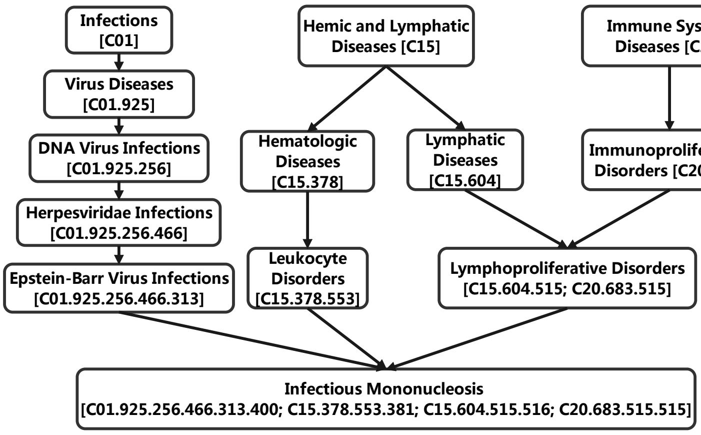

Infections [C01]
Virus Diseases [C01.925]
DNA Virus Infections [C01.925.256]
Herpesviridae Infections [C01.925.256.466]
Epstein-Barr Virus Infections [C01.925.256.466.313]
Hemic and Lymphatic Diseases [C15]
Hematologic Diseases [C15.378]
Lymphatic Diseases [C15.604]
Leukocyte Disorders [C15.378.553]
Immune System Diseases [C20]
Immunoproliferative Disorders [C20.683]
Lymphoproliferative Disorders [C15.604.515; C20.683.515]
Infectious Mononucleosis [C01.925.256.466.313.400; C15.378.553.381; C15.604.515.516; C20.683.515.515]

Figure 2. The directed acyclic graph(DAG) of infectious mononucleosis.

2.3. Graph Embedding

A network is a structure composed of nodes and edges, where nodes are connected by edges. Graph embedding is a distributed representation of network structure, which includes node embedding, edge embedding and subgraph embedding. Node embedding can map the discrete nodes in the network to a continuous vector space, and each node has an unique vector representation. Node embedding data after encoding the topological information of the network are very useful inputs relative to machine learning algorithms for downstream tasks, such as node classification and link prediction. In this section, we describe the ideas of Node2vec [28] and GATNE [22], which are used in our prediction as modules.

2.3.1. Node2vec

Node2vec is a graph embedding method, which obtains the nearest neighbor sequence of nodes by a biased random walk. Two hyperparameters p and q are introduced to balance the depth first search (DFS) and breadth first search (BFS). For the miRNA–miRNA similarity network, given a current miRNA node v, let t denote its previous miRNA node. The probability of accessing to the next miRNA x is calculated as follows:

$$P(c_{i+1} = x | c_i = v) = \begin{cases} \frac{\alpha_{pq}(t,x) * w_{vx}}{Z}, & \text{if} \quad (v,x) \in E \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

$$\alpha_{pq}(t,x) = \begin{cases} \frac{1}{p}, & \text{if} \quad d_{tx} = 0 \\ 1, & \text{if} \quad d_{tx} = 1 \\ \frac{1}{q}, & \text{if} \quad d_{tx} = 2 \end{cases} \tag{5}$$

where $\alpha_{pq}(t,x)$ and w_{vx} are the transition probability and similarity between miRNA v and miRNA x, respectively; Z is the normalized constant. Based on the hypothesis that similar miRNAs are more likely to be associated with the same disease and vice versa, we set $p = 1, q = 2$ in this study. Assume that f is the mapping function of the miRNA node u to the embedding vector. $N_s(u)$ is defined as the set of adjacent nodes of u, which is sampled through the sampling strategy S. The optimization goal of node2vec is to maximize the co-occurrence probability of the nearby miRNAs for a given current miRNA node [28].

$$max_f \sum_{u \in V} \log Pr(N_s(u) | f(u)). \tag{6}$$

After embedding, the higher the similarity between two miRNA nodes, the closer the Euclidean distance of their features. The same operation is adopted in disease-disease similarity network.

2.3.2. GATNE

For an attributed network $G(V, E, A)$, $V = \{v_1, v_2, ..., v_n\}$, $E = \{e_{ij}|v_i, v_j \in V\}$ and $A = \{x_i|v_i \in V\}$ are the sets of nodes, edges and node features, respectively. If the number of types of the nodes and edges in G is larger than 1, G is called an attributed multi-layer heterogeneous network (AMHN). $G_r = \{V, E_r, A\}$ is denoted as a subnetwork of edge type r. GATNE [22] is an inductive embedding algorithm on AMHN, composed of three parts: base embedding, edge embedding and node attributes. The core idea of GATNE is to aggregate neighbors from different layers to the current node and then to generate different vector representations for the nodes on each edge type.

Base embedding of a node is shared in different layers. For miRNA nodes in attributed multi-layer miRNA-disease heterogeneous network, the base embedding is obtained through the attributes of miRNAs, and it can be calculated as follows:

$$b_i = h_z(x_i), \tag{7}$$

where x_i is the feature vector of miRNA v_i and h_z is the transformation function for miRNAs for which its type is defined as z.

The edge embedding of attributed multi-layer miRNA-disease heterogeneous network on each relation type is initialized by the transformation function that takes node attributes as input, then the idea of neighbor aggregation in GraphSAGE [29] is adopted to aggregate edge embedding on different levels in single layer. For miRNA node on relation type r, we have the following:

$$d_{i,r}^{(0)} = g_{z,r}(x_i), \tag{8}$$

$$d_{i,r}^{(k)} = aggregator(\{d_{i,r}^{(k-1)}, \forall v_j \in N_{i,r}\}), \tag{9}$$

where $g_{z,r}$ is the transformation function, k is defined as the kth level neighbor of miRNA v_i and $N_{i,r}$ is the set of neighbor nodes of v_i. To take into account the interplays between different relation types, the self-attention mechanism [30] is performed on the concatenate representation constructed by the edge embeddings of nodes from different layers:

$$D_i = (d_{i,1}, d_{i,2}, ..., d_{i,m}), \tag{10}$$

$$a_{i,r} = softmax(w_r^T \tanh(W_r D_i))^T, \tag{11}$$

$$u_{i,r} = \alpha_r M_r D_r a_{i,r}, \tag{12}$$

where m is the number of relation types and w_r, W_r and M_r are parameters of relation type r, which can be trained by parameter optimization.

The embedding representation of miRNA node v_i on relation type r is calculated as follows.

$$v_{i,r} = \beta_r D_z^T x_i + b_i + u_{i,r}. \tag{13}$$

The same embedding operations are adopted on disease nodes.

In parameter optimization, the meta-path-based random walk [31] and skip gram [32] on the heterogeneous network and negative sampling are adopted to optimize the parameters of GATNE. The objective function is as follows:

$$\begin{aligned} E &= -logP_\theta(\{v_j|v_j \in C\}|v_i) \\ &= -log\sigma(c_j^T \cdot v_{i,j}) - \sum_{l=1}^{L} E_{v_k \sim P_t(v)}[log\sigma(-c_k^T \cdot v_{i,r})], \end{aligned} \tag{14}$$

where the context C of path $P = (v_{p1,...,pt})$ is generated by meta-path-based random walk and c_k is the context embedding of node v_k, which is a negative sample randomly selected from distribution $P_t(v)$, $\sigma(x) = 1/(1 + exp(-x))$. θ represents all parameters in GATNE.

2.4. *mDLinker*

In this paper, we propose a novel algorithm (named mDLinker) to predict multiple association types between miRNAs and diseases through a multi-layer heterogeneous network embedding technique. The first step of mDLinker is to calculate miRNA–miRNA similarities based on the miRNAs' sequences and calculate disease–disease similarities based on the DAG structure of the diseases. Denote the miRNA–miRNA similarity network as $G_M = (V_M, E_M)$ with the weighted adjacency matrix $A_M \in R^{m\times m}$, where $A_M(i,j)$ is the similarity between miRNAs i and j, $V_M = \{1, 2, ..., m\}$ is the set of miRNA nodes and E_M is the set of edges. The second step of mDLinker is to obtain the feature matrix of miRNA nodes $X_M \in R^{m\times NE}$ by Node2vec, where NE represents the feature dimension of miRNA nodes after embedding. Similarly, we obtained the disease–disease similarity network $G_D = (V_D, E_D)$ with the weighted adjacency matrix $A_D \in R^{n\times n}$, $V_d = \{1, 2, ..., n\}$ and determine the feature matrix $X_D \in R^{n\times NE}$ through Node2vec. Then, we constructed an attributed multi-layer miRNA-disease heterogeneous network (AMH-MD), where the only difference between each layer of AMH-MD is the type of association. One of these layers is shown in Figure 3. In AMH-MD, the topologies of all single miRNA-disease association and the relationships between different types of association, as well as the feature of nodes, are well described.

Let MD-P denote an miRNA-disease pair that is known with an association, but the association is not specified with any type; and let MD-T denote an miRNA-disease-type triple, meaning the type of the association between the miRNA and disease is specified. Given the complicated correlations between the MD-Ts, for each association type, the graph embedding technique GATNE is applied by our method mDLinker to aggregate topology information from different layers to the current layer such that the embedding representation of the miRNAs and diseases on each layer can be characterized. Then, we train a machine learning classifier on these experimentally confirmed MD-Ts and a set of selected negative samples, where the feature vectors of MD-Ts are obtained by concatenating the two vector representations of corresponding diseases and miRNAs. The algorithm flow diagram is shown in Figure 4.

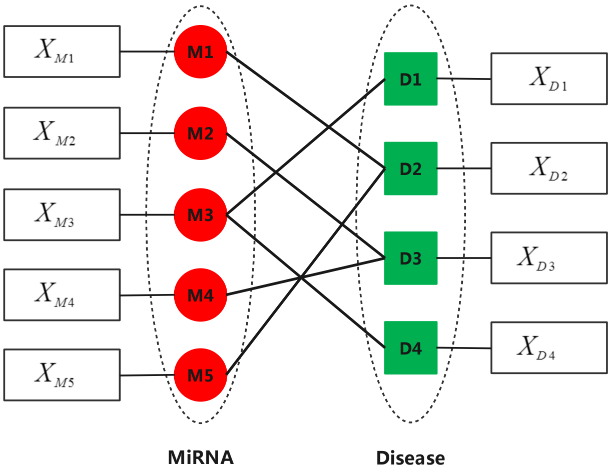


Figure 3. A single layer in our attributed multi-layer miRNA-disease heterogeneous network, where X_{Mi} and X_{Dj} represent the feature vector of miRNA Mi and that of disease Dj, respectively.

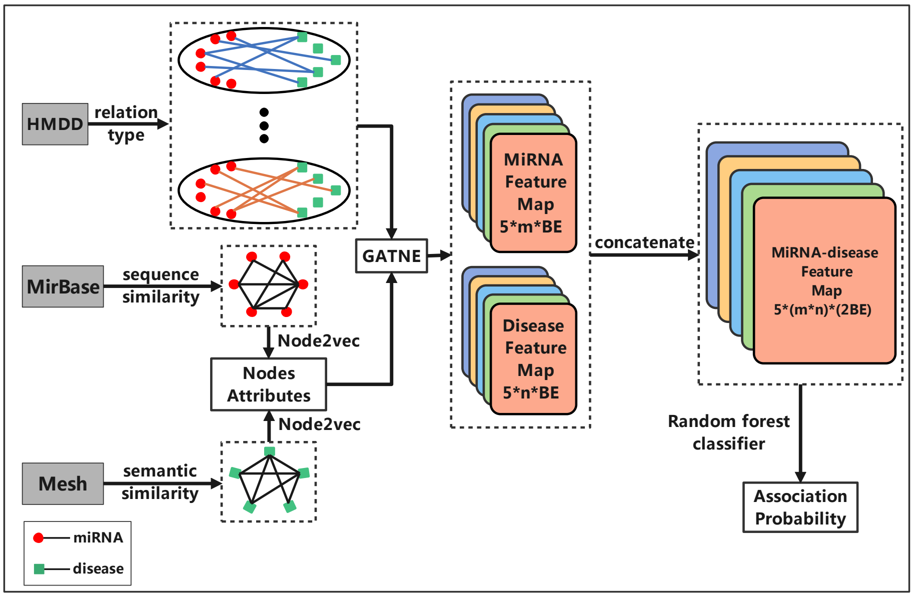


Figure 4. An illustration of predicting multiple types of miRNA-disease associations by mDLinker. There are m miRNAs, n diseases and five layers in the network representing the five types of associations. BE is the dimension number of the miRNA vector representation and also the dimension number of the disease vector representation embedded by GANTE.

3. Experiments and Results

3.1. Experimental Setting

In order to validate the performance of mDLinker, we adopted two kinds of experiments, termed CV-Type and CV-Triple, as similarly used by Huang et al. [21] for a fair comparison.

By CV-Type, we randomly divided the confirmed MD-P instances into 10 equal parts, one of which was reserved as a test set and the others are used as the training set. In both the test set and training set, experimentally verified miRNA-disease-type triples are regarded as the positive samples, while the unconfirmed or non-existent MD-Ts are served as negative samples. For each MD-P in the test set, the association probabilities of all types of MD-T are predicted. If the top one ranked MD-T is confirmed by HMDD, the corresponding MD-P is predicted correctly. The purpose of the CV-Type is to demonstrate the ability of exploring the most reliable MD-T from the corresponding MD-P. Precision, recall and f1 of top one are calculated in CV-Type.

In CV-Triple, we randomly divided all MD-T instances into 10 equal parts, one part reserved as a test set, and the others form the training set. Due to the lack of MD-T negative samples provided by biologists, we constructed a set N_{sample} of negative samples from the unconfirmed or non-existent MD-T by a strategy similar to [33,34]. For relation type r, we calculated an average representation feature $f_{avg,r}$ of all positive samples in the training set. The Euclidean distances from $f_{avg,r}$ to all unconfirmed or non-existent MD-T are measured, where the average Euclidean distance is denoted as dis_r. For MD-T i, if the distance $dis_{r,i}$ from $f_{avg,r}$ to i is greater than dis_r, MD-T i can be regarded as a more reliable negative sample on relation type r. In this experiment, we randomly selected negative samples equal to the number of positive samples on each edge type from N_{sample}. Three performance indicators of AUPR, AUC and $f1$-measure are calculated to confirm the effectiveness of mDLinker.

3.2. Performance by Different Classifiers

We investigated the performance of different classifiers on predicting the potential miRNA-disease-type triples under the setting CV-Type. Several classifiers were attempted, including decision tree, naive Bayes, logistic regression, KNN, SVM and random forest.

The performances are shown in Table 1. As an ensemble classifier, random forest achieved the best performance.

Table 1. Ten-fold cross validation performance in CV-Type by different classifiers based on HMDD v3.2.

Method	Top-1 Precision	Top-1 Recall	Top-1 F1
Decision tree	0.4203	0.3224	0.3649
Naive Bayes	0.4748	0.3640	0.4121
Logistic Regression	0.4960	0.3803	0.4305
KNN	0.6001	0.4602	0.5209
SVM	0.5706	0.4357	0.4952
Random Forest	0.6509	0.4991	0.5649

3.3. Performance Comparison with the State of the Art

Several methods have been proposed to predict the types of miRNA-disease associations. We compared the performance of our mDLinker with two newest algorithms TFAI [21] and TDRC [21] on the same data set used in our paper from HMDD v3.2. NLPMMDA [20] can effectively predict each association type, respectively, but it only models the binary association on each type of miRNA-disease association. On the one hand, according to the comparison results of Huang et al. [21], TDRC algorithm has the best performance compared to the other methods, and it is significantly superior to NLPMMDA. On the other hand, we cannot run NLPMMDA on our data set because the authors of NLPMMDA did not provide the source code of it. Thus, we did not perform the comparison between mDLinker and NLPMMDA in our study. The parameters in TFAI and TDRC are set as suggested by the original paper. From Tables 2 and 3, one can observe that mDLinker consistently outperforms TFAI and TDRC under all evaluation criteria. By the ranking-based evaluation (Table 2), mDLinker, TFAI and TDRC achieved top one f1 of 0.5649, 0.4832 and 0.5071, respectively. This suggests that utilizing non-linear relationships between miRNA nodes and disease nodes has greatly contributed to the high performance of identifying potential MD-T from MD-P. Under the setting CV-Triple, the most significant improvement is made again by mDLinker. In particular, the scores of F1, AUPR and AUC are improved by 4.99%, 5.15% and 6.41%, respectively, in comparison with the previous best model TDRC.

Table 2. Performance comparison among TFAI, TDRC and mDLinker under the setting CV-Type.

Method	Top-1 Precision	Top-1 Recall	Top-1 F1
TFAI	0.5874	0.4501	0.4832
TDRC	0.6116	0.4686	0.5071
mDLinker	0.6509	0.4991	0.5649

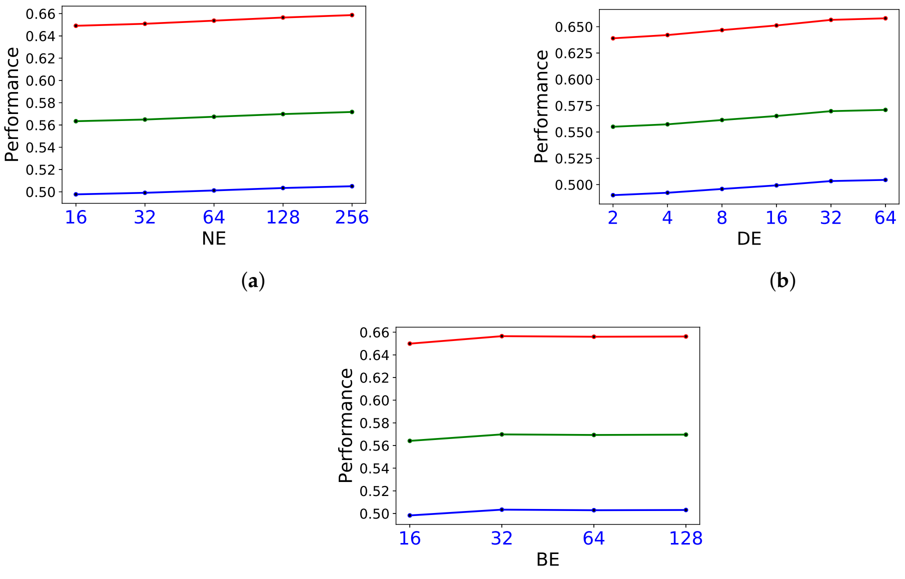
Table 3. Performance comparison among TFAI, TDRC and mDLinker under the setting CV-Triple.

Method	AUPR	AUC	F1
TFAI	0.9261	0.912	0.8559
TDRC	0.93	0.9222	0.865
mDLinker	0.9799	0.9737	0.9291

3.4. Sensitivity Study on Parameters

Sensitivity analyses were conducted on three hyper-parameters: NE (the attribute dimension of miRNA nodes and disease nodes), DE (the dimension of edge embedding) and BE (the dimension of base embedding). These three hyper-parameters determine the framework of mDLinker and affect the performance of prediction.

Figure 5 shows the performance when different hyper-parameter settings were tested under CV-Type. NE was ranged in $\{16, 32, 64, 128, 256\}$. It can be observed that when NE becomes larger, the performance trends better. Therefore, we set $NE = 128$ to prevent overfitting. DE and BE are searched in the ranges $\{2, 4, 8, 16, 32, 64\}$ and $\{16, 32, 64, 128\}$. The best and most stable performance is achieved when $DE = 32$ and $BE = 32$.



(**a**)

(**b**)

(**c**)

Figure 5. Sensitivity analysis on different parameter settings, including dimension of node attribute (**a**), dimension of edge embedding (**b**) and dimension of base embedding (**c**), where the red polyline represents Top-1 precision, the green polyline represents Top-1 F1 and the blue polyline represents Top-1 recall.

3.5. Case Study: miRNA-Disease Association Types Predicted beyond the HMDD Databases

The database HMDD v3.2 has recorded much more miRNA-disease associations than its early version 2.0. We conducted a case study to observe whether the newly recorded miRNA-disease associations and their types in HMDD v3.2 can be predicted by a mDLinker model established from HMDD v2.0. For this case study, we used the disease name mapping table provided by Huang et al. [35] to map disease names from HMDD v2.0 to HMDD v3.2, and we selected 15 unique disease names in HMDD v3.2 that each have at least 20 miRNA-disease-type triples (i.e., at least 20 miRNA-disease associations of different types) as test cases. Node2vec and GATNE are sampled via biased random walk and the meta-path-based random walk, respectively. In order to eliminate the error caused by sampling, we ran it 50 times and considered the top 20 MD-Ts with the highest consensus frequency as the prediction result. Figure 6 shows the detailed prediction performance by mDLinker on the 15 diseases. We can observe that many of the predicted associations are actual associations recorded in HMDD v3.2, especially for lung neoplasms and hepatocellular carcinoma, where 85% of the top 20 predicted associations are actual associations confirmed in HMDD v3.2.

Our second case study is to use all the associations and the association type information stored at HMDD v3.2 to train our prediction model and then to apply this model to predict currently unknown miRNA-disease associations and their types. We run the model 50 times. Each time, from the miRNA-disease-type triples that are not recorded in HMDD v3.2, we stored those triples with the highest association probability predicted by our model. Over 50 times, the top 10 predicted associations with the most consensus in the stored top lists were regarded as the most likely miRNA-disease associations with a specific type. For example, an miRNA-target association type between mir-224 and Gastric Neoplasms predicted by our method mDLinker was confirmed by Fang et al. [36], and mir-

224-5p negatively regulates OPCML in gastric cancer tissues; the abnormal expression of miR-148a in prostate cancer predicted by MDLinker is consistent with the findings by Fujita et al. [37] that the expression level of miR-148a in PC3 and DU145 hormone-resistant prostate cancer cells is lower than that in PrEC normal prostate epithelial cells. More details of these predicted associations are presented in Table 4. Although none of them are currently recorded in HMDD v3.2, all of these predictions can be confirmed by the literature works from NCBI (see the PMIDs at the fourth column of the table).

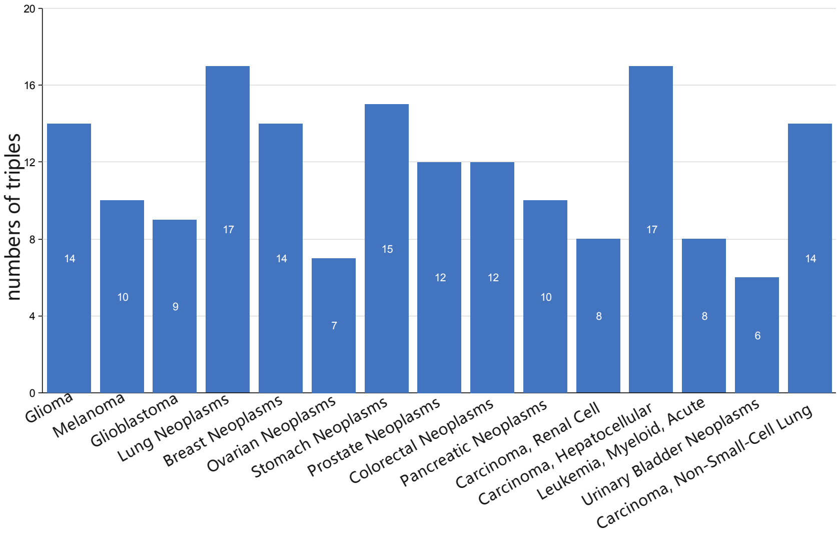


Figure 6. Case study on 15 diseases unique in HMDD v3.2. The training model is based on the miRNA-disease-type triples recorded in HMDD v2.0. Numbers of predicted miRNA-disease associations, which are confirmed by HMDD v3.2, are highlighted in the middle of the bars.

Table 4. miRNA-disease associations predicted by our model trained on HMDD v3.2 with their association types and their supporting literature from NCBI.

MiRNA	Disease	Type	PMID	Experimental Methods	Description
hsa-mir-224	Gastric Neoplasms	Target	32359894	RT-qPCR; Western blot analysis	OPCML is negatively regulated by miR-224-5p in Gastric cancer tissues.
hsa-mir-193a	Carcinoma, Hepatocellular	Target	30710422	qRT-PCR; Dual luciferase reporter assay; RNA pull-down assay	miR-193a-5p inhibits the growth of hepatocellular carcinoma by targeting SPOCK1.
hsa-mir-31	Gastric Neoplasms	Target	30677405	Silico analysis; Dual luciferase reporter assay	Zeste homolog 2 (ZH2) is the potential target of miR-31 in AGS cells to inhibit Gastric cancer.

hsa-mir-218-1	Carcinoma, Hepatocellular	Target	30003726	Fluorescence protein analysis;
				RT-qPCR;
				Western blotting
					miR-218 suppresses the growth of hepatocellular carcinoma by inhibiting the expression of proto-oncogene Bmi-1.

hsa-mir-148a	Prostate Neoplasms	Tissue	20406806	the trypan blue; dye exclusion assay	miR-148a expression levels are lower in PC3 and DU145 hormone-refractory prostate cancer cells than PrEC normal human prostate epithelial cells.

hsa-mir-218-1	Breast Neoplasms	Target	29378184	RT-qPCR analysis;
				Luciferase reporter assay;
				Cancer biostatistical analysis
					miR-218 regulates breast cancer progression by targeting Lamins.

hsa-let-7	Carcinoma, Hepatocellular	Target	27821157	MTT assay;
				western blot;
				immunofluorescence;
				luciferase-reporter assay
					Let-7 inhibits the self-renewal of stem cell-like cells by regulating Wnt signaling pathwayand EMT.

Table 4. *Cont.*

MiRNA	Disease	Type	PMID	Experimental Methods	Description
hsa-mir-193a	Colorectal Carcinoma	Target	29104111	qRT-PCR; Western bolt analysis	MiR-193a-3p plays a tumor suppressive role by targeting KRAS in colorectal adenocarcinoma patients.
hsa-mir-200c	Carcinoma, Cervical	Target	27693631	Luciferase reporter; qRT-PCR assays	Disrupting MALAT1/miR-200c sponge decreases invasion and migration in endometrioid endometrial carcinoma.

hsa-mir-34c	Ovarian Neoplasms	Target	32308421	qRT-PCR; MTT;
				Western blot assays;
				Immunoprecipitation;
				Flow cytometry analysis
					miR-34c targets MET to improve the Anti-Tumor effect of Cisplatin on ovarian cancer.

4. Conclusions

It is well known that there is a strong correlation between disease and miRNA. miRNA can be used as a diagnostic marker and therapeutic target for disease. Unless evidence of miRNA-disease association is confirmed, it is not sufficient to understand pathology at the molecular level of miRNA. In this study, we have proposed an innovative algorithm (mDLinker) based on embedding techniques and machine learning ideas in order to predict the heterogeneous types of miRNA-disease associations and showed the significance and usefulness of utilizing non-linear relationships in uncovering potential miRNA-disease-type triples. By mDLinker, the miRNAs and diseases attributes are obtained through Node2vec to embed the miRNA sequence information and DAG structure information of the diseases, respectively. All the nodes in AMH-MD are projected into a vector on each association by the powerful embedding technique GATNE. In order to calculate the probability of each association type between an miRNA and a disease, the random forest classifier is adopted. Compared with the state-of-the-art algorithms, mDLinker achieves significantly better performance in the systematical experiments. Furthermore, it performs excellent in the prediction of currently unknown associations in the HMDD v3.2, as demonstrated in our case studies. The mDLinker is shown to be a reliable model in exploring multiple relationships between miRNAs and diseases, and it is a useful tool in enhancing understanding of the molecular basis of disease formation.

Author Contributions: D.-L.Y., Z.-G.Y. and J.L. conceived the experiment; D.-L.Y. conducted the experiment; D.-L.Y., Z.-G.Y., G.-S.H. and J.L. analyzed the results; D.-L.Y., Z.-G.Y. and J.L. wrote the manuscript; J.L., V.A. and G.-S.H. reviewed the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: The authors thank the anonymous reviewers for their valuable suggestions. This work was supported by funds from the National Natural Science Foundation of China (grant numbers: 11871061 and 12026213); Collaborative Research project for Overseas Scholars (including Hong Kong and Macau) of National Natural Science Foundation of China (Grant No. 61828203); The National Key Research and Development Program of China (grant number: 2020YFC0832405); and Innovation Foundation of Qian Xuesen Laboratory of Space Technology; Hunan Provincial Innovation Foundation for Postgraduate (grant number: CX20200611).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: The authors thank the anonymous reviewers for their valuable suggestions.

Conflicts of Interest: There are no competing interests.

Appendix A

Table A1. Details of the two data sets in this work: *a* represents miRNA-disease pair association, and *b* represents miRNA-disease-type triple.

HMDD	MiRNA	Disease	MD-P [a]	MD-T [b]				
				Type-1	Type-2	Type-3	Type-4	Type-5
v2.0	321	168	1506	355	215	441	676	0
v3.2	695	445	12,495	1578	519	3300	5822	5079

References

1. Lynam-Lennon, N.; Maher, S.G.; Reynolds, J.V. The roles of microRNA in cancer and apoptosis. *Biol. Rev. Camb. Philos. Soc.* **2009**, *84*, 55–71. [CrossRef]
2. Garzon, R.; Marcucci, G.; Croce, C.M. Targeting microRNAs in cancer: Rationale, strategies and challenges. *Nat. Rev. Drug Discov.* **2010**, *9*, 775–789. [CrossRef]
3. Li, C.; Feng, Y.; Coukos, G.; Zhang, L. Therapeutic microRNA strategies in human cancer. *AAPS J.* **2009**, *11*, 747–757. [CrossRef]
4. 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]
5. Li, Y.; Qiu, C.; Tu, J.; Geng, B.; Yang, J.; Jiang, T.; Cui, Q. HMDD v2.0: A database for experimentally supported human microRNA and disease associations. *Nucleic Acids Res.* **2014**, *42*, D1070–D1074. [CrossRef] [PubMed]
6. Hu, Y.; Liu, C.M.; Qi, L.; He, T.Z.; Shi-Guo, L.; Hao, C.J.; Ma, X. Two common SNPs in pri-miR-125a alter the mature miRNA expression and associate with recurrent pregnancy loss in a Han-Chinese population. *RNA Biol.* **2011**, *8*, 861–872. [CrossRef] [PubMed]
7. Cimmino, A.; Calin, G.A.; Fabbri, M.; Iorio, M.V.; Ferracin, M.; Shimizu, M.; Wojcik, S.E.; Aqeilan, R.I.; Zupo, S.; Dono, M.; et al. miR-15 and miR-16 induce apoptosis by targeting BCL2. *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 13944–13949. [CrossRef] [PubMed]
8. Li, C.; Lu, S.; Wang, Y.; Guo, S.; Zhao, T.; Wang, X.; Song, B. Influence of microRNA 34a on proliferation, invasion and metastasis of HCT116 cells. *Mol. Med. Rep.* **2017**, *15*, 833–838. [CrossRef] [PubMed]
9. Bailey, M.H.; Tokheim, C.; Porta-Pardo, E.; Sengupta, S.; Bertrand, D.; Weerasinghe, A.; Schein, J.E. Comprehensive characterization of cancer driver genes and mutations. *Cell* **2018**, *174*, 1034–1035. [CrossRef]
10. Yao, Q.; Chen, Y.; Zhou, X. The roles of microRNAs in epigenetic regulation. *Curr. Opin. Chem. Biol.* **2019**, *51*, 11–17. [CrossRef]
11. Mitchell, P.S.; Parkin, R.K.; Kroh, E.M.; Fritz, B.R.; Wyman, S.K.; Pogosova-Agadjanyan, E.L.; Tewari, M. Circulating microRNAs as stable blood-based markers for cancer detection. *Proc. Natl. Acad. Sci. USA* **2008**, *105*, 10513–10518. [CrossRef]
12. Wu, Y.; Li, Q.; Zhang, R.; Dai, X.; Chen, W.; Xing, D. Circulating microRNAs: Biomarkers of disease. *Clin. Chim. Acta* **2021**, *516*, 46–54. [CrossRef]
13. Zhang, Y.; Liu, D.; Chen, X.; Li, J.; Li, L.; Bian, Z.; Zhang, C.Y. Secreted monocytic miR-150 enhances targeted endothelial cell migration. *Mol. Cell* **2010**, *39*, 133–144. [CrossRef]
14. Rao, Y.M.; Shi, H.R.; Ji, M.; Chen, C.H. MiR-106a targets Mcl-1 to suppress cisplatin resistance of ovarian cancer A2780 cells. *J. Huazhong Univ. Sci. Technol. Med. Sci.* **2013**, *33*, 567–572. [CrossRef]
15. Zu, L.; Xue, Y.; Wang, J.; Fu, Y.; Wang, X; Xiao, G.; Wang, J. The feedback loop between miR-124 and TGF-β pathway plays a significant role in non-small cell lung cancer metastasis. *Carcinogenesis* **2016**, *37*, 333–343. [CrossRef]
16. 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.* **2019**, *47*, D1013–D1017. [CrossRef]
17. Tan, X.; Zhou, L.; Wang, H.; Yang, Y.; Sun, Y.; Wang, Z.; Li, H. Differential expression profiles of microRNAs in highly and weakly invasive/metastatic pancreatic cancer cells. *Oncol. Lett.* **2018**, *16*, 6026–6038. [CrossRef] [PubMed]
18. Cui, C.; Cui, Q. The relationship of human tissue microRNAs with those from body fluids. *Sci. Rep.* **2020**, *10*, 5644. [CrossRef]
19. Chen, X.; Yan, C.C.; Zhang, X.; Li, Z.; Deng, L.; Zhang, Y.; Dai, Q. RBMMMDA: Predicting multiple types of disease-microRNA associations. *Sci. Rep.* **2015**, *5*, 13877. [CrossRef] [PubMed]
20. Zhang, X.; Yin, J.; Zhang, X. A Semi-Supervised Learning Algorithm for Predicting Four Types MiRNA-Disease Associations by Mutual Information in a Heterogeneous Network. *Genes* **2018**, *9*, 139. [CrossRef] [PubMed]
21. Huang, F.; Yue, X.; Xiong, Z.; Yu, Z.; Liu, S.; Zhang, W. Tensor decomposition with relational constraints for predicting multiple types of microRNA-disease associations. *Brief. Bioinform.* **2020**, *22*, bbaa140.
22. Cen, Y.K.; Zou, X.; Zhang, J.W.; Yang, H.X.; Zhou, J.G.; Tang, J. Representation Learning for Attributed Multiplex Heterogeneous Network. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD 2019), Anchorage, AK, USA, 4–8 August 2019; pp. 1358–1368.
23. 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]
24. Li, H.Y.; You, Z.H.; Wang, L.; Yan, X.; Li, Z.W. DF-MDA: An effective diffusion-based computational model for predicting miRNA-disease association. *Mol. Ther.* **2021** , *29*, 1501–1511. [CrossRef] [PubMed]

25. 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] [PubMed]
26. Che, K.; Guo, M.; Wang, C.; Liu, X.; Chen, X. Predicting MiRNA-Disease Association by Latent Feature Extraction with Positive Samples. *Genes* **2019**, *10*, 80. [CrossRef]
27. Zheng, K.; You, Z.H.; Wang, L.; Zhou, Y.; Li, L.P.; Li, Z.W. DBMDA: A Unified Embedding for Sequence-Based miRNA Similarity Measure with Applications to Predict and Validate miRNA-Disease Associations. *Mol. Ther. Nucleic Acids* **2020**, *19* , 602–611. [CrossRef] [PubMed]
28. Grover, A.; Leskovec, J. Node2vec: Scalable Feature Learning for Networks. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2016), San Francisco, CA, USA, 13–17 August 2016; pp. 855–864.
29. Hamilton, W.; Ying, Z.T.; Leskovec, J. Inductive representation learning on large graphs. In Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017; pp. 1024–1034.
30. Lin, Z.H.; Feng, M.W.; Santos, C.N.; Yu, M.; Xiang, B.; Zhou, B.; Bengio, Y. A Structured Self-attentive Sentence Embedding. In Proceedings of the International Conference on Learning Representations, Toulon, France, 24–26 April 2017.
31. Dong, Y.X.; Chawla, N.V.; Swami, A. Metapath2vec: Scalable Representation Learning for Heterogeneous Networks. In Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2017), Halifax, NS, Canada, 13–17 August 2017; pp. 135–144.
32. Mikolov, T.; Chen, K.; Corrado, G.S.; Dean, J. Efficient Estimation of Word Representations in Vector Space. In Proceedings of the Workshop at ICLR, Scottsdale, AZ, USA, 2–4 May 2013.
33. Li, C.; Liu, H.; Hu, Q.; Que, J.; Yao, J. A Novel Computational Model for Predicting microRNA-Disease Associations Based on Heterogeneous Graph Convolutional Networks. *Cells* **2019**, *8*, 977. [CrossRef]
34. Luo, P.; Li, Y.; Wu, F.X. Enhancing the prediction of disease-gene associations with multimodal deep learning. *Bioinformatics* **2019**, *35*, 3735–3742. [CrossRef] [PubMed]
35. Huang, Z.; Liu, L.; Gao Y.; Shi, J.; Cui, Q.; Li, J.; Zhou, Y. Benchmark of computational methods for predicting microRNA-disease associations. *Genome Biol.* **2019**, *20*, 202. [CrossRef] [PubMed]
36. Fang, X.; Dong, Y.; Yang, R.; Wei L. LINC00619 restricts gastric cancer progression by preventing microRNA-224-5p-mediated inhibition of OPCML. *Arch. Biochem. Biophys.* **2020**, *689*, 108390. [CrossRef]
37. Kojima, K.; Fujita, Y.; Nozawa, Y.; Deguchi, T.; Ito, M. MiR-148a attenuates paclitaxel resistance of hormone-refractory, drug-resistant prostate cancer PC3 cells by regulating MSK1 expression. *J. Biol. Chem.* **2010**, *285*, 19076–19084.