BMC Bioinformatics

RESEARCH

Open Access

MHESMMR: a multilevel model for predicting the regulation of miRNAs expression by small molecules

Yong-Jian Guan[1], Chang-Qing Yu[1*], Li-Ping Li[2,4*], Zhu-Hong You[3], Meng-meng Wei[1], Xin-Fei Wang[1], Chen Yang[1] and Lu-Xiang Guo[1]

*Correspondence:
xaycq@163.com;
cs2bioinformatics@gmail.com

[1] School of Information Engineering, Xijing University, Xi'an, China
[2] College of Grassland and Environment Sciences, Xinjiang Agricultural University, Urumqi, China
[3] School of Computer Science, North-Western Polytechnical University, Xi'an, China
[4] College of Agriculture and Forestry, Longdong University, Qingyang, China

Abstract

According to the expression of miRNA in pathological processes, miRNAs can be divided into oncogenes or tumor suppressors. Prediction of the regulation relations between miRNAs and small molecules (SMs) becomes a vital goal for miRNA-target therapy. But traditional biological approaches are laborious and expensive. Thus, there is an urgent need to develop a computational model. In this study, we proposed a computational model to predict whether the regulatory relationship between miRNAs and SMs is up-regulated or down-regulated. Specifically, we first use the Large-scale Information Network Embedding (LINE) algorithm to construct the node features from the self-similarity networks, then use the General Attributed Multiplex Heterogeneous Network Embedding (GATNE) algorithm to extract the topological information from the attribute network, and finally utilize the Light Gradient Boosting Machine (LightGBM) algorithm to predict the regulatory relationship between miRNAs and SMs. In the fivefold cross-validation experiment, the average accuracies of the proposed model on the SM2miR dataset reached 79.59% and 80.37% for up-regulation pairs and down-regulation pairs, respectively. In addition, we compared our model with another published model. Moreover, in the case study for 5-FU, 7 of 10 candidate miRNAs are confirmed by related literature. Therefore, we believe that our model can promote the research of miRNA-targeted therapy.

Keywords: LINE, microRNA, Small molecule, Generally attributed multiplex heterogeneous network embedding, Machine learning

Introduction

As an emerging biomarker for medical and diagnostics, microRNA (miRNA) is a small single-stranded endogenously-initiated non-coding RNA molecule [1]. Since Ambros et al. discovered the first miRNA *lin-4*, about 28,000 miRNA molecules have been found in animals, plants and some viruses [2, 3]. Previously, the genomic structure and subtypes of protein, such as transcription factors and epigenetic mediators, were regarded as the only regulators of gene expression. However, researchers reveal the critical role of miRNAs in post-transcriptional regulatory mechanisms. Mature miRNA can bind to the

3'-untranslated region end of target mRNA, which triggers a decrease in the expression level of specific DNA [4]. This also suggests that miRNA expression levels affect multiple cellular functions, such as embryonic development, regulating substance metabolism, mediating signal transduction, cell division and apoptosis [5, 6]. In the human body, over 60% of transcription is regulated by miRNAs [7]. Since each miRNA can regulate the expression of many genes, each miRNA can regulate multiple cellular signalling pathways at the same time [8].

Cell activity is inseparable from the post-transcriptional regulation of miRNA. Meanwhile, many research papers indicated that the dysregulation of miRNA is related to disease occurrence, most notably cancer. Whether over-expression of carcinogenic miRNAs (oncomiRs) or down-regulation of tumor suppressor miRNAs (TSmiRs) may cause malignant tumours [9, 10]. Thus, miRNAs can be regarded as a biomarker for diagnosis [11]. People conducted a kind of medical treatment strategy based on the miRNA, called miRNA-target therapeutics [12, 13]. Its main modality is to regulate the expression level of oncomiRs or TSmiRs through SM. Since the special tertiary structure of miRNA, SM can bind to miRNA with high affinity and specificity. For example, Naro et al. discovered the first SM inhibitor of miRNA for suppressing the expression of miR-21 by the luciferase-base screening of more than 300,000 small molecules [14]. Miravirsen, a kind of oligonucleotide-based miR-122 inhibitor, has entered clinical trials and is well tolerated in non-human primates, which greatly reduces the burden of HCV and liver cancer [15]. Chandrasekhar et al. identified that aza-Flavanones could be an inhibitor of miR-4644, which was helpful to arrest and eliminate human breast tumor cells [16]. Besides, for a long time, it is extensively supposed that only proteins can be used as drug targets. But in fact, only about 600 kinds of disease modification proteins can be targeted by drugs. miRNA-targeted drugs are an important supplement to the pharmaceutical industry [17, 18]. In summary, discovering the regulation relation between miRNAs and small molecules harbours major implications for advancing miRNA-target therapeutics and drug development.

So far, the methods for discovering miRNA-target SM drugs can be divided into three categories. The first category is the high-throughput screening approach which uses high-throughput screening techniques to identify SM inhibitors or activators of miRNAs. For example, Zhang et al. presented a method based on miRNA 3D structure to discover miRNA-target SM which can regulate miRNA activity [19]. They utilized MC-fold to obtain miRNA structure. Similar to using Auto Dock to calculate the affinities between binding sites and ligand, they computed RNA-compatible score of SMs by molecule docking-based high-throughput screening techniques. Another category of approaches considers the structure of RNA base sequence. The most famous case is the web server of Inforna developed by Disney *et at.*, which predicts the association between SM-miRNA through motif alignment on a large scale in the databases [20]. The third category of the method is based on fluorescence detection assays. Bose et al. proposed a new method for identifying SM targeting miRNA in vitro using a molecule beacon [21]. The oligonucleotide hybridization probes are labelled with a fluorophore and a quencher when the beacon binds to the target miRNA. These studies have been instrumental in developing novel miRNA targeting SM drugs and old SM drug repositioning. Anyways, detecting the regulation of miRNAs expression by SMs through biological experiments

is time-consuming and labor-intensive because the Bio-data is diverse and voluminous. Therefore, researchers intensified studies into developing computational methods to predict the association between SMs and miRNAs, hoping to narrow down the candidate drug searching scope and accelerate the process of drug development.

In recent years, a series of diverse computational models have been proposed to predict the association between miRNAs and SMs [22]. These miRNA-SM association prediction methods can be divided into two categories. The first category is sequence similartiy-based methods. For example, Lv et al. constructed an integrated SM-miRNA association network that combines the miRNA self-similarity network, SM self-similarity network and the known SM to miRNA targeting relationship network [23]. And they performed the improved random walk with restart algorithm (RWR) on the integrated SM-miRNA network, which allowed the random walk to learn samples on the various layers of the network. Finally, they ranked miRNA by the relevance score to each SMs, thus screening for potential miRNA targeting SMs. Jiang et al. leveraged the functional similarity of gene expression profiles under drug treatment and miRNA perturbation for SM-miRNA association prediction [24]. Meng et al. proposed the predicting model RWNS based on a three layers network including miRNA, SMs and diseases [25]. They considered multiple functional similarities such as SM chemical structure similarity, disease phenotype-based similarity and miRNA targeted gene functional consistency-based similarity. The integrated multiple types of functional similarities were constructed in a three layers network and implemented the random walk algorithm on the network. Deepthi et al. conducted a method to predict the relation between SM drugs and miRNA via the convolutional neural network (CNN). The miRNA similarity network and the SM similarity network were used as the features of miRNA and SM. The principal component analysis was implemented to reduce the dimensions of features and the CNN model was trained to extract the high-order information. Finally, they used the support vector machines for identifying the potential relation between miRNAs and SMs. Besides, Guan et al. developed the SM-miRNA association prediction model called the GISMMA model with the graphlet interaction-based inference [26]. The graphlet interaction aimed at describing the complex relationship between the miRNA similarity network and the SM similarity network. By counting the number of 28 types of graphlet interaction isomers, the GISMMA model can yield the predicted score of the potential relation between miRNA and SM. The second category is heterogeneous network-based methods. Li et al. presented the SMiR-NBI model to find miRNAs that can be the potential biomarkers for anticancer drugs. They constructed the SMiR-NBI model by a network-based inference. Specifically, they first initialized the resource scores of miRNAs based on the SM-miRNA adjacent matrix. Then the resource of miRNA was averagely distributed among the SM drugs that were directly linked to that miRNA in the network. Similarly, the SM drugs redistributed the resources to adjacent miRNAs after they integrated the resource from adjacent miRNAs. The final resource score of each miRNA represents the probability that it can be used as the biomarker for a certain anti-cancer drug.Wang et al. presented an approach of a triple layer heterogeneous network (TLHNSMA) to predict the association between SMs and miRNAs [27]. They exploited the functional similarities and relationships of miRNAs, SMs and diseases to construct a triple layers network. Then they developed an interactive updating algorithm

to propagate the information across the three layers heterogeneous network. Anyways, there are three major disadvantages of these methods. First, most of the previous methods can only predict whether the SM can interact with miRNA but ought not to predict the regulation relation of the SM to the miRNA. These methods are unable to satisfy drug development and target selection because miRNAs may function as oncomiRs or TSmiRs. Thus, the key to advancing the research progress of miRNA-targeted therapy is to identify the SM modulators that inhibit oncomiRs and activate TSmiRs. Second, since most methods rely on the functional similarity of miRNA and SMs, these methods are constrained by complex side information. Therefore, there is a urgent need of an efficient and accurate auxiliary tool for the prediction of the SM regulation with miRNA.

One of the challenges in predicting the association between miRNA and SM is to identify whether their regulatory relationship is up-regulated or down-regulated. To address this challenge, we were inspired by the successful application of the attributed multi-layer heterogeneous network for predictions of multi-typed associations between miRNAs and diseases [28]. In this study, for predicting the miRNA-SM regulation relation, we introduced the attributed multi-layer heterogeneous network containing miRNA self-similarity and SM self-similarity. And we proposed a novel multilevel model called MHESMMR. The multilevel mdoel is composed of attributed multi-layer heterogeneous network and networks embedding methods. In detail, our proposed model consists of three steps. First, we carry out the LINE algorithm on the miRNA self-similarity and SM self-similarity for generating node features and then utilizes these node features to construct the attributed multi-layer heterogeneous network of miRNAs and SMs. And then, the GATNE algorithm is used for learning the representation features from the attributed multi-layer heterogeneous network. Finally, we feed these features into the LightGBM classifier to identify the probable SM modulators. To evaluate the performance of the proposed model, we predict the SM2miR under fivefold cross-validation. Furthermore, we compared the proposed model with other node feature extraction methods and machine learning classifiers, and the experiment results prove that the proposed model is a robust and efficient auxiliary tool for screening SM modulators for miRNA.

Materials and methods

Dataset

In the experiment, we collected the data about the regulation relation between SMs and miRNAs to evaluate the performance of the proposed model from the latest version of the SM2miR database [29]. The SM2miR database is a manually curated database that collected numerous SM's effects on miRNA expression validated by the previous literature. According to the expression patterns of miRNA, the SM2miR database was divided into two parts, up-regulated pairs and down-regulated pairs, which correspond to Dataset 1 and Dataset 2, respectively. After pre-processing steps, we obtained 541 miRNA, 831 SM drugs and 2377 miRNA-SM pairs. Among these, 1394 up-regulation pairs belong to Dataset 1 and 983 down-regulation pairs belong to Dataset 2. The known SM-miRNA regulation relation pairs were regarded as positive samples.

In general, we describe a bipartite heterogeneous network of SM-miRNA regulation relations in which SM drugs and miRNAs are represented by nodes, and the relationships between them are represented by edges. The imbalanced problems may introduce bias

into the experiment results. Thus, the same number of positive samples should be selected from unlabelled samples to generate the negative samples. In theory, the unlabelled samples selected in this manner may involve some potential SM-miRNA relation pairs. To do so, we carry out a negative sample selecting method based on the sequence proximity as similarly used by Yu et al. for negative sampling [30]. In terms of SM drugs, we generate MACCS fingerprints from SMILES to represent the SM drug chemical structure by the "RDKit" python package [31, 32]. To measure the proximity between each SM drug, we calculate the value of Tanimoto coefficients, a quantitative way for sequence alignment, based on their MACCS fingerprint.

Then, the regulation relations between any SMs and any miRNAs was computed. For example, we suppose that the regulation relation between miRNA1 and SM1 is unknown but miRNA1 can be inhibited by SM2, SM3 and SM4. The regulation relations between miRNA1 and SM1 can be calculated as follow:

$$r_{SM1}^{miRNA1} = \frac{\bar{s}_{SM2} + \bar{s}_{SM3} + \bar{s}_{SM4}}{3} = \bar{s} \tag{1}$$

where $\bar{s}$ denotes the mean value of Tanimoto coefficients of SM1-SM2, SM1-SM3, and SM1-SM4. We computed all of the regulation relations for unlabelled SM-miRNA pairs in the same way. Only the pairs of regulation relations score less than 0.1 were selected as the negative samples. Finally, we selected 1394 negative samples for Dataset 1 and 983 negative samples for Datset2.

Node attributes of heterogeneous network by graph embedding

Graph embedding methods allow distributed representation of network structure, which can be divided into three categories including node embedding, edge embedding and substructure embedding. The node representation maps the nodes to the embedding space and each node can be represented by a vector. By doing this, the node embedding data containing the topological information of the graph are very effective inputs relative to the machine learning model for downstream classification tasks.

The LINE is a graph embedding method based on neighbourhood similarity assumptions proposed by Tang et al. and it is suitable for a weighted network [33]. In a complex network, if two vertices are direct neighbours, they are considered to have first-order proximity. On the other hand, if there are multiple first-order proximity vertices between two nodes, they are considered to have second-order proximity. From these two aspects, the main idea of the LIEN algorithm can be divided into two parts.

First-order proximity is to describe the local similarity in the graph. And the LINE with first-order proximity can only be applied to the undirected graph. The joint probability p_1 between two vertices v_i and v_j on the edge $e(i, j)$ can be defined as:

$$p_1(v_i, v_j) = \frac{1}{1 + \exp(-\vec{u}_i^T \vec{u}_j)} \tag{2}$$

where $\vec{u}_i$ and $\vec{u}_j$ are the low-dimensional the low-dimensional representation vectors of v_i and v_j. It can describe the relationship between vertices from the perspective of embedding space. The distribution $p(*, *)$ over the space $V \times V$ is defined as Formula (2). And its empirical probability $\hat{p}_1$ can be defined as:

$$\hat{p}_1(i,j) = \frac{w_{ij}}{W} \tag{3}$$

$$W = \sum_{(i,j)\in E} w_{ij} \tag{4}$$

where w_{ij} denotes the weight of the edge between vertices v_i and v_j, and W denotes the sum of all weights of the edges. The goal of our optimization formula is to minimize the difference between p_1 and $\hat{p}_1$, so the objective function is defined as follows:

$$O_1 = d(p1(*,*), \hat{p}1(*,*)) \tag{5}$$

where $d()$ represents the function used to measure the difference between two kinds of distributions. And the Kullback–Leibler (KL) divergence can be introduced to the above formula to replace the $d(*,*)$. The final optimized formula is defined as:

$$O_1 = -\sum_{(i,j)\in E} w_{ij} log p_1(v_i, v_j) \tag{6}$$

Thus, all of the vertices can be represented as $\{\vec{u}_i\}_{i=1...|V|}$ in the d-dimensional space by optimizing the objective function.

The LINE also considers the second-order proximity between vertices. And the LINE with second-order proximity can be applied on both directed and undirected graphs. For a directed edge $e(i,j)$, the probability that vertex v_i and vertex v_j are directly connected can be defined as:

$$p_2 = (v_j|v_i) = \frac{\exp(\vec{u}_j^{'T} \cdot \vec{u}_i)}{\sum_{k=1}^{|V|} \exp(\vec{u}_k^{'T} \cdot \vec{u}_i)} \tag{7}$$

where $|V|$ denotes the number of vertices in the graph. And the empirical distribution is defined as:

$$\hat{p}_2(v_j|v_i) = \frac{w_{ij}}{d_i} \tag{8}$$

where d_i denotes the out-degree of v_i and w_{ij} denotes the weight of the edge $e(i,j)$. In order to make the low-dimensional representation of the conditional distribution of context $p_2(\cdot|v_i)$ as close as possible to the empirical distribution $\hat{p}_2(\cdot|v_i)$, the objective function can be defined as:

$$O_2 = \sum_{i\in V} \alpha_i d(\hat{p}_2(*,*), p_2(*,*)) \tag{9}$$

where α_i denotes the prestige of the vertex v_i and set as the degree of the vertex v_i in this study. As mentioned above, $d(*,*)$ is replaced by KL-divergence. Thus, the final optimization function is defined as:

$$O_2 = \sum_{(i,j)\in E} w_{ij} log p_2(v_j|v_i) \tag{10}$$

Finally, each vertex can be represented by a d-dimensional vector $\vec{u}_i$ by finding $\{\vec{u}_i\}_{i=1...|V|}$ after minimizing the objective function. We applied the LINE algorithm to the miRNA self-similarity network calculated by the Tanimoto Coefficient. After graph embedding, if the properties of the two miRNAs are very similar, the embedding vectors between them will also be very close. We also performed the same operation on the SM self-similar network.

Attributed multiplex heterogeneous network embedding

With the development of graph embedding, or network representation learning, exploring non-linear properties are critically important in extracting topological information from heterogeneous networks. There is an emerging graph embedding technology, called general attributed multiplex heterogeneous network embedding (GATEN). The GATNE algorithm aims to integrate the attribute features of the nodes and the multiple relationships between different types of nodes. Furthermore, it can project the information of nodes and non-linear relationships in the network into a relatively low-dimensional representation vector. Figure 1 shows the GATNE algorithm in inductive mode.

We assume that a relationship graph G with a set of vertices $V = \{v_1, v_2, ..., v_n\}$, a set of edges $E = \{e_{ij}|v_i, v_j \in V\}$ and node attributed features $A = \{x_i|v_i \in V\}$, that is $G = \{V, E, A\}$. If the vertices and edges are of more than one type, G is a multi-layer heterogeneous network that represents as $G_r = (V, E_r, A)$ and r denotes the types of relationships between two vertices. In general, the GATNE aggregates neighbour information and attributes information from the inductive context to the current vertices and generates feature vectors for each vertex at different layers. The GATNE is an inductive learning model with the combination of two parts: base embedding and edge embedding.

The base embedding of vertex v_i is shared in different types of edges. The based embedding b_i is calculated by a transform function defined as follow:

$$b_i = h_z(x_i) \tag{11}$$

where h_z is a transformation function of attribute feature x_i of vertex v_i and the corresponding vertex type is represented by z.

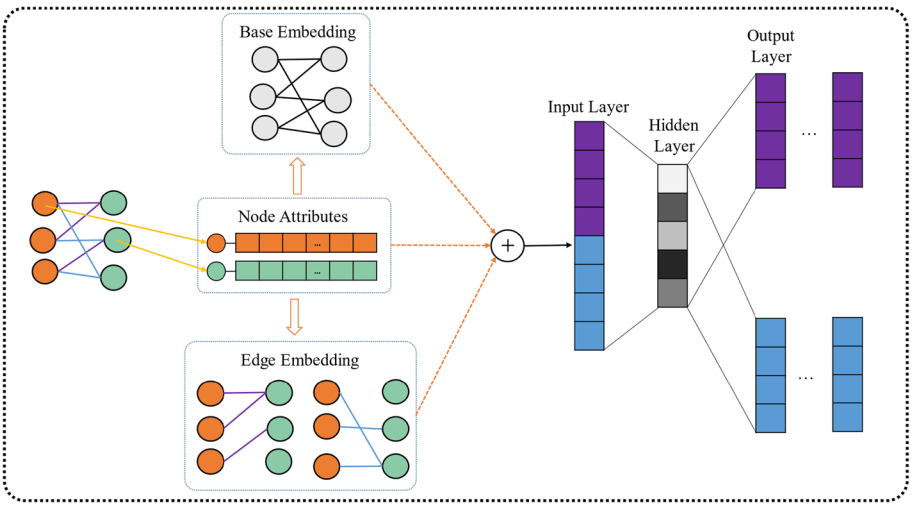


Fig. 1 Illusion of GATNE in inductive mode

In the edge embedding, the initial edge embedding $u_{i,r}^{(0)}$ for vertices is constructed by the transformation function with vertices attribute features A. as input. GraphSAGE is a graph neural network technology based on information aggregation [34]. The GATNE draws from the neighbour aggregator of the GraphSAGE to aggregate edge embedding vectors of vertex v_i on layer r. The initial edge embedding and the mean aggregator function are as following:

$$u_{i,r}^{(0)} = g_{z,r}(x_i) \tag{12}$$

$$u_{i,r}^{(k)} = aggregator\left(\{u_{i,r}^{(k-1)}, \forall v_j \in N_{i,r}\}\right) \tag{13}$$

where the transformation function of z type vertex v_i in relation r is denoted as $g_{z,r}$. $u_{i,r}^{(K)}$ denoted the K-th level edge embedding after aggregation and $N_{i,r}$ represent the neighbour of vertex v_i in relation r. Then all edge embedding $u_{i,r}$ in relation r of vertex v_i are concatenated as U_i with size s-by-m, where s represents the dimension of edge embeddings:

$$U_i = (u_{i,1}, u_{i,2}, ..., u_{i,m}) \tag{14}$$

The self-attention mechanism is performed on the U_i to calculate the coefficients $c_{i,r} \in R^m$ of linear combination of edge embedding in U_i on relation type of r, the function is formula as:

$$c_{i,r} = softmax(w_r^T \tanh(W_r U_i))^T \tag{15}$$

where w_r and W_r are the trainable parameters of relation type r and trained by optimization framework.

In general, the embedding representation vector of miRNAs and SM molecules on relation type r are computed by the jointly optimization function as follow:

$$v_{i,r} = b_i + \alpha_r M_r^T U_i a_{i,r} + \beta_r D_Z^T x_i \tag{16}$$

where b_i is the based embedding of vertex v_i.α_r is the hyper-parameter indicating the proportion of edge embedding in the entire embedding. And $M_r \in R^{s \times d}$ is trainable transformation matrix.

In parameter optimization framework, the GATNE integrated base embedding and edge embedding by the random walk and skip gram model on the attributed multilayer heterogeneous network [35, 36]. Except random walk, meta-path-based methods are also commonly used in research in the field of bioinformatics in recent years. Meta-paths can be used to mine similarities and influences among network nodes. Based on these meta-paths, the similarity or weight between different nodes can be calculated to obtain more accurate recommendation results. At the same time, new relationships can also be discovered through meta-paths to improve the diversity and innovation of prediction models [37, 38].The meta-path-based random walk is used to generate vertices sequences to learn embedding. In detail, we suppose a graph $G_r = (V, E_r, A)$ and a meta-path scheme $T : V_1 \rightarrow V_2 \rightarrow ...V_t... \rightarrow V_l$, where l is the length of the meta-path scheme. And the transition probability of random walk is defined as:

$$p(v_j|v_i,T)=\begin{cases}\frac{1}{|N_{i,r}\cap v_{t+1}|} & (v_i,v_j)\in E_r, v_j\in V_{t+1}\\ 0 & (v_i,v_j)\in E_r, v_j\notin V_{t+1}\\ 0 & (v_i,v_j)\notin E_r\end{cases} \tag{17}$$

where $v_i \in V_t$ and $N_{i,r}$ is the neighbourhood of vertices v_i in relation type r.. The meta-path-based random walk aims at digging out the semantic relationship between two different types of vertices for integrating by the skip-gram model. Finally, the objective function is defined as:

$$\begin{aligned}E &= -\log P_\theta(\{v_j|v_j\in C\}|v_i)\\ &= -\log\sigma(c_j^T\cdot v_{i,r})-\sum_{l=1}^{L}E_{vk\sim P_t(v)}[\log\sigma(-c_k^T\cdot v_{i,r})]\end{aligned} \tag{18}$$

where C is the context of vertex v_i in the path $P=(v_1,...v_l)$ and c_k is the embedding of vertex v_i. σ represents the sigmoid function and L is the number of negative samples equal to positive samples. Among v_k is randomly drawn from the distribution $Pt(v)$ which defined on the set of corresponding vertices v_i.

LightGBM

In this study, we introduce a maching learning method as the classifier. LightGBM is a type of machine learning algorithm based on Gradient Boosting Decision Tree (GBDT) [39]. It is an efficient and fast gradient boosting framework developed by Microsoft. The lightGBM algorithm contains two novel techniques, namely Gradient-based One-Side Sampling (GOSS) and the Exclusive Feature Bundling (EFB), which can handle a large number of data instances and a large number of data features without overfitting problem, respectively [40]. LightGBM uses a histogram-based decision tree algorithm to discretize continuous features into discrete histogram features, thereby reducing data storage space and computational complexity. LightGBM uses a growth strategy called leaf-wise. The leaf-wise growth strategy selects the current optimal leaf node for splitting each time, which can quickly find the direction in which the loss function decreases the fastest, thus speeding up the training of the model.

MHESMMR

In this work, owing to effective application of network embedding techniques in the bioinformatic field in the post-genomic era, we propose a novel computaional method named MHESMMR to predict multiple regulatory relations between miRNAs and SMs. MHESMMR can be describe in following five steps: (1) use the dataset to construct a multi-layer heterogeneous network, (2) construct the self-similarity networks of SM and miRNA by Tanimoto coeffcient, (3) generating node features by using LINE algorithm on the miRNA self-similarity and SM self-similarity network, (4) apply GATNE algorithm to aggregate the behavior information from the attributed multi-layer heterogeneous network for learning representation features(5) identify the probable SM modulators by the machine learning classifier, where the feature vectors of miRNA-SM are obtained by concatenating two representation features of corresponding miRNAs and SMs. The flowchart of the MHESMMR model is shown in Fig. 2.

Experimental results and discussion

Performance evaluation criterion

To validate the performance of the proposed model, we implemented a series of evaluation criteria. And fivefold cross-validation is adopted to ensure the rigor of the experiment. In detail, the positive samples and negative samples are equally divided into 5 folds. In each round of fivefold cross-validation, one of the folds is used as a testing sample set so that the prediction scores can be used using the proposed method. These prediction scores can reflect the possibility that an SM drug can regulate the expression of a miRNA. In our performance evaluation, if the positive sample in the test set has a high predictive score and the negative sample has a low predictive score, this indicates that the proposed model has good performance.

Moreover, we monitored accuracy (Acc.), sensitivity (Sen), specificity (Spec.) and Matthews Correlation Coefficient (MCC) to comprehensively evaluate the proposed model as follows:

$$Acc. = \frac{TN + TP}{TN + TP + FN + FP} \tag{19}$$

$$Sen. = \frac{TP}{FP + FN} \tag{20}$$

$$Spec. = \frac{TN}{TN + FP} \tag{21}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{22}$$

where TP is the number of positive samples that prediction score is higher than the threshold; FN is the number of positive samples that prediction score is lower than the threshold; FP is the number of negative samples that prediction score is higher than

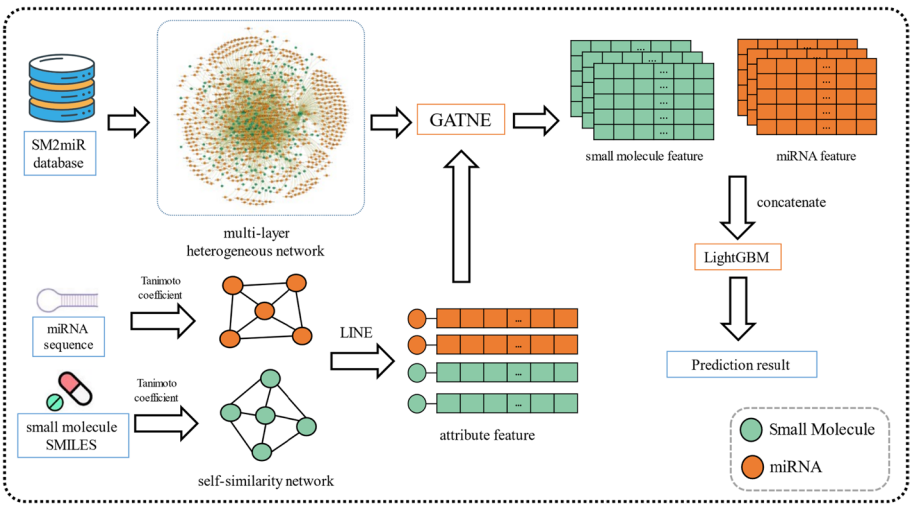


Fig. 2 Framework of the MHESMMR model to predict miRNA-SM regulatory relations

the threshold; TN is the number of negative samples that prediction score is lower the threshold, respectively. To show the results more intuitively, we drew the receiver operating characteristic (ROC) curves and precision-recall (PR) curves. The area under the ROC curve (AUC) and area under PR (AUPR) were also used for the evaluation of model performance [41, 42]. If the value of AUC is 0.5 that denotes a purely random prediction and 1 denotes a perfect prediction.

Sensitivity analysis on parameters

To obtain the best prediction performance, we performed the sensitivity analysis on the base embedding dimension and the edge embedding dimension. In this part, the sensitivity analysis was conducted on two hyper-parameters of the GATNE algorithm. Figure 3 illustrates the line chart of average AUC values, which was generated by the LightGBM classifier and influenced by the features dimension and the edge embedding dimension. It can be observed that when the base dimension is set to 128, the best results are obtained on the two data sets. The proposed model gets the best results when the edge embedding dimension of data set 1 and data set 2 is 64 and 32 respectively.

In addition, an additional experiment was carried out to prove the effectiveness of the node attribute features generated by the self-similarity networks. Specifically, we removed the node attribute features and utilized the transductive mode to generate node features just relying on the network structure. As expected, without any node attribute feature, the MHESMMR model yielded average AUCs of 0.937 and 0.9509 on Dataset1 and Dataset2, which is lower than that obtained with attribute feature inputs. Figure 4 displays the ROC curve for this experiment. These results prove the combination of the attribute feature and the graph topology feature can improve the prediction performance.

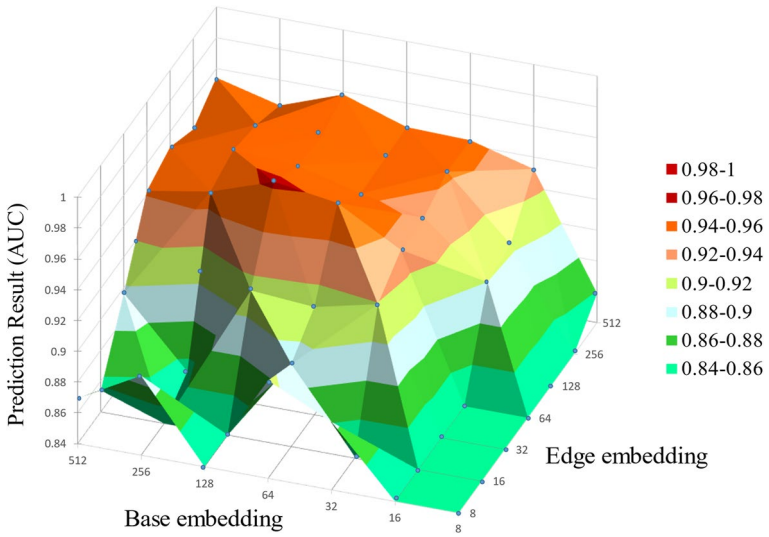


Fig. 3 Sensitivity analysis on the dimension of base embedding and edge embedding

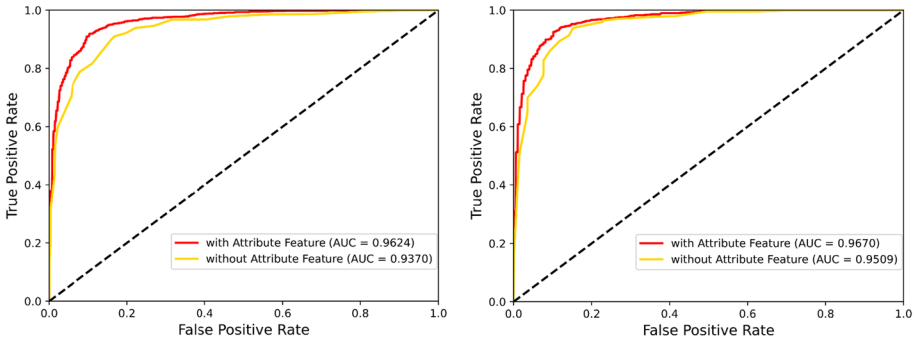


Fig. 4 Difference of prediction performance using MHESMMR model with/without attribute feature input on Dataset1 (**a**) and Dataset2 (**b**)

Assessment of prediction ability

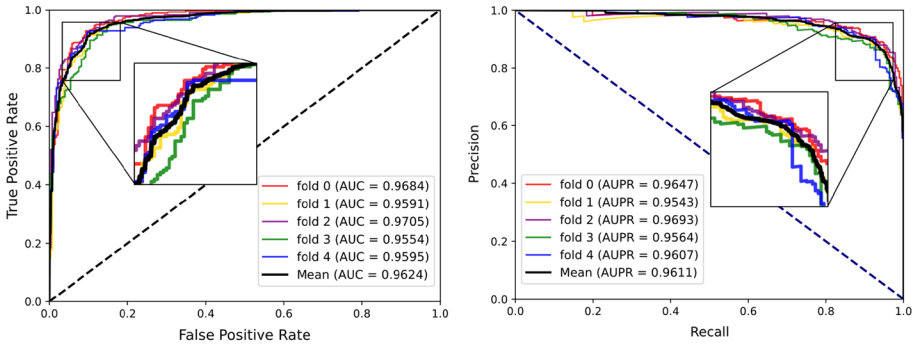
To evaluate the prediction ability of the MHESMMR model while avoiding overfitting, we conducted fivefold cross-validation experiment on two datasets for our proposed model. To maintain consistency, all parameters of these experiments were consistent in this study. In Dataset1, we achieved the average results of Acc., Prec., Sen., MCC, AUC and AUPR of 90.55%, 92.73%, 89.25%, 82.10% 0.9624, 0.9607 and the standard deviations of 0.91%, 1.65%, 2.03%, 1.8%, 0.0065, 0.0050, respectively. In Dataset2, we obtained the average evaluation criteria of 90.97%, 92.74%, 85.28%, 79.98%, 0.9622, 0.9605 and the standard deviations of 1.51%, 1.76%, 3.25%, 2.94%, 0.0099, 0.1102, respectively. The results of the proposed model are summarized in the Table 1 and 2 when adopting the fivefold cross-validation on two datasets. The ROC and PR curves of the fivefold cross-validation experiment are shown in Figs. 5 and 6. All these results indicated a reliable predictive ability of our model.

Table 1 fivefold cross-validation performance for Dataset1

Fold	Acc. (%)	Sen. (%)	Spec. (%)	MCC (%)	AUC	AUPR
0	91.17	94.95	87.41	82.58	0.9684	0.9647
1	90.31	90.29	90.32	80.61	0.9591	0.9543
2	91.20	92.81	89.61	82.45	0.9705	0.9693
3	89.05	92.81	85.30	78.32	0.9554	0.9564
4	91.02	92.81	89.25	82.10	0.9595	0.9607
Average	90.55 ± 0.91	92.73 ± 1.65	88.38 ± 2.03	81.21 ± 1.80	0.9624 ± 0.0065	0.9611 ± 0.0050

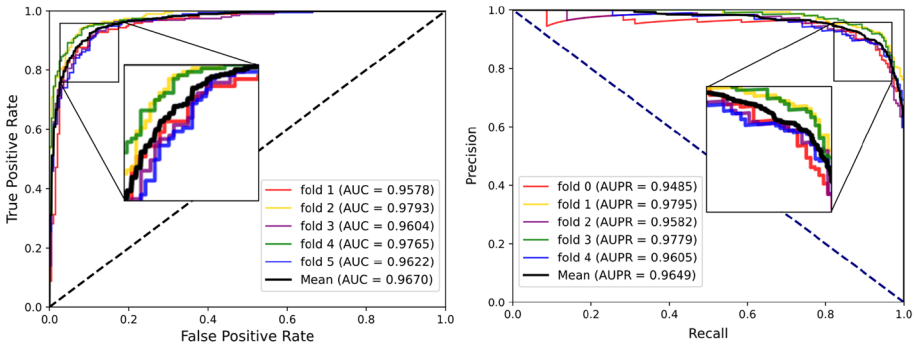
Table 2 fivefold cross-validation performance for Dataset2

Fold	Acc. (%)	Sen. (%)	Spec. (%)	MCC (%)	AUC	AUPR
0	89.77	92.82	86.73	79.69	0.9578	0.9485
1	92.62	91.84	93.4	85.25	0.9793	0.9795
2	90.03	90.26	89.8	80.05	0.9604	0.9582
3	92.62	94.39	90.86	85.3	0.9765	0.9779
4	89.82	94.39	85.28	79.98	0.9622	0.9605
Average	90.97 ± 1.51	92.74 ± 1.76	89.21 ± 3.25	82.05 ± 2.94	0.9670 ± 0.0099	0.9649 ± 0.1102



(a) (b)

Fig. 5 ROC curve (**a**) and PR curve (**b**) performed by MHESMMR on Dataset 1



(a) (b)

Fig. 6 ROC curve (**a**) and PR curve (**b**) performed by MHESMMR on Dataset 2

Ablation experiments

In MHESMMR model, the feature construction can be dived into two modules: node attribute feature construction and graph embedding feature construction. In ablation experiments, we verify which parts of the model contribute the most to the final performance. We constructed two prediction models using only one kinds of feature construction method. The first model only uses the GATNE algorithm to construct features, namely MHESMMR(G), which initial features of the attributed heterogeneous network are set to unit vectors. The second model is called MHESMMR(A), in which the extracted node features are directly input into the classifier to obtain prediction results. To ensure the fairness of the experiment, the same parameters and data set were used in all experiments. The experimental results were objectively recorded in Table 3. For the convenience of comparison, Fig. 7 was used to describe the comparison between the data in the ablation experiment and the original data. Figure 7 shows that the best prediction results can be achieved by combining the two models. Among them, MHESMMR(G) has a better prediction effect than MHESMMR(A), which proves that GATNE algorithm makes a greater contribution to the overall model.

Table 3 Ablation experiment result on Dataset1 and Dataset2

Model	Dataset	Acc. (%)	Sen. (%)	Spec. (%)	MCC (%)
MHESMMR	Dataset1	90.55 ± 0.91	92.73 ± 1.65	88.38 ± 2.03	81.21 ± 1.80
	Dataset2	90.97 ± 1.51	92.74 ± 1.76	89.21 ± 3.25	82.05 ± 2.94
MHESMMR(G)	Dataset1	87.11 ± 1.26	88.50 ± 2.39	76.09 ± 2.93	78.31 ± 2.41
	Dataset2	89.63 ± 3.22	90.73 ± 1.75	85.91 ± 3.71	81.61 ± 1.96
MHESMMR(A)	Dataset1	81.36 ± 1.54	84.35 ± 2.96	73.31 ± 3.69	75.88 ± 3.65
	Dataset2	88.43 ± 2.97	89.25 ± 3.01	86.31 ± 4.02	79.36 ± 2.45

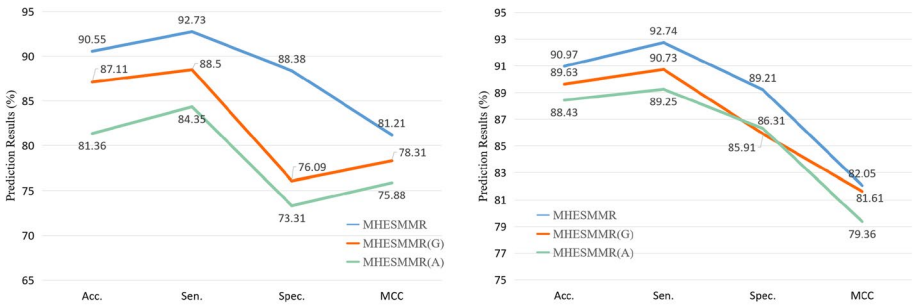


Fig. 7 Ablation experiment result on Dataset1 and Dataset2

Performance by different classifiers

Machine learning algorithms are widely used in molecular interaction prediction models [43]. In order to prove the superiority of our classification strategy, we selected a number of classic algorithms commonly used in the field of bioinformatics to replace our classification method and compare the results.In the experiment, we used several popular machine learning algorithms to construct the prediction model including Logistic Regression (LR), Navi Bayes (NB), Support Vector Machines (SVM), Random Forest (RF) and LightGBM [44–48]. The performance of models based on Dataset1 from five-fold cross-validation is shown in Figs. 8 and 9.

When predicting miRNA-SM regulation relation for the Dataset1, we yielded average Acc., Sen., Spec., MCC, AUC and AUPR values of 90.55%, 92.73%,88.38%, 81.21%, 0.9594 and 0.9611S with corresponding standard deviations of 0.91, 1.65, 2.03, 1.80, 0.0065 and 0.0046, respectively. When predicting miRNA-SM regulation relation for the Dataset2, we yielded average Acc., Sen., Spec. and MCC values of 90.55%, 92.73%,88.38%, 81.21% with corresponding standard deviations of 0.91, 1.65, 2.03, 1.80, respectively. From these results, we can note that, among these five different prediction models, the LightGBM-based model achieved the highest Acc. on Dataset1 and Dataset2 of 90.55% and 90.97%. Moreover, the RF-based model yielded the second-highest Acc. Of 85.70% and 86.95%. Finally, the LightGBM model was selected for constructing the predicting model.

Method comparison

To my knowledge, the only computational model that can predict two types of regulatory relationship (up-regulated or down-regulated) between miRNAs and SMs is

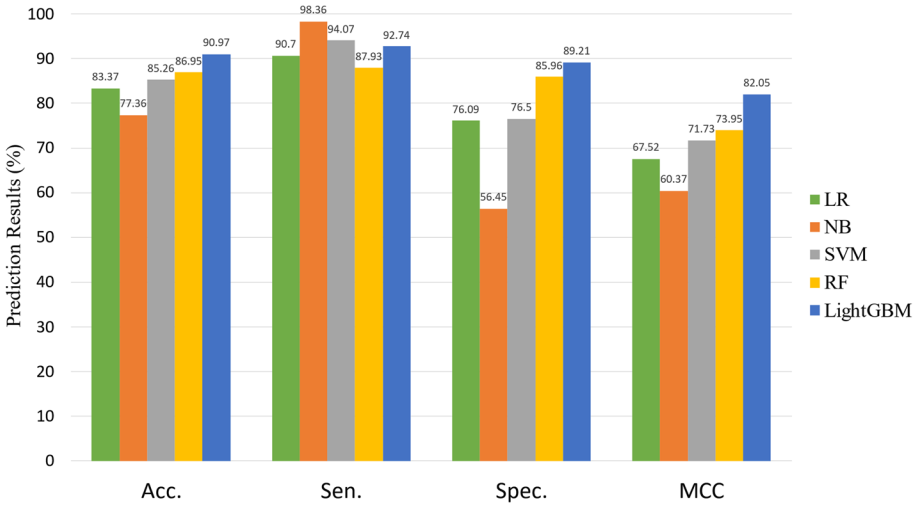


Fig. 8 Experimental result of different classifiers on Dataset1

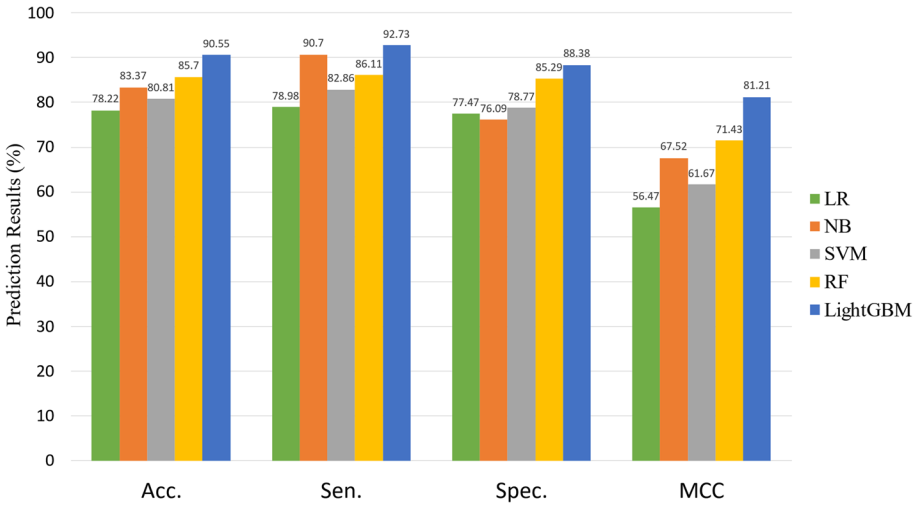


Fig. 9 Experimental result of different classifiers on Dataset2

called PSRR [49]. To further demonstrate the predictive ability of the MHESMMR model, we compared it with the PSRR model on Dataset1 and Dataset2. The PSRR model constructs miRNA attribute features by 2-mer and 4-mer based on miRNA base sequences. And the166-dimensional MACCS fingerprints are calculated as the descriptors of SMs according to the SMILES of SMs. Finally, the PSRR model concatenates the two kinds of features and uses RF for classification prediction. In addition, we applied several previous models for predicting the interaction relationship between miRNA and small molecule drugs to our dataset, including BNEMDI, GCLAMS and ELDMA. BNEMDI used BiNE algorithm to extract tological feature in bipartite graph and predict miRNA-small molecule associaiton by DNN model. GCLAMS is a prediciton model based on heterogenous graph fusion neural netwok. ELDMA utilized the integrated pairwise similarities of small molecule and miRNA and convolutional neural network to extract intricate features. To show the results

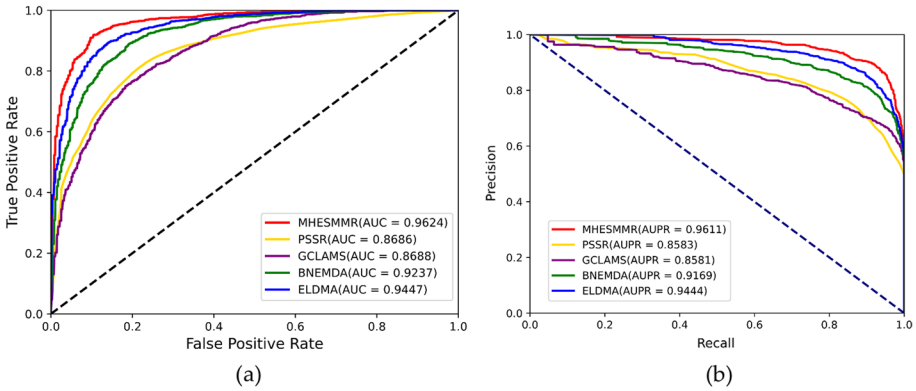


Fig. 10 Prediction performance of models for SM-miRNA up-regulation pairs (Dataset1)

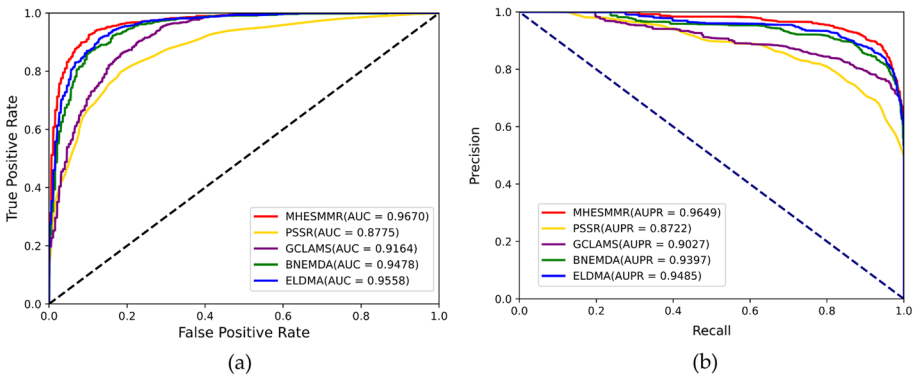


Fig. 11 Prediction performance of models for SM-miRNA down-regulation pairs (Dataset2)

Table 4 Comparison of experimental results of MHESMMR and PSSR

Model	Dataset	Acc. (%)	Sen. (%)	Spec. (%)	MCC (%)	AUC	AUPR
BNEMDI	Dataset1	84.69 ± 2.07	86.75 ± 2.97	82.64 ± 2.43	73.29 ± 2.09	0.9237 ± 0.0364	0.9169 ± 0.0145
	Dataset2	88.83 ± 0.55	90.29 ± 1.61	87.38 ± 2.31	77.74 ± 1.83	0.9478 ± 0.0045	0.9397 ± 0.0051
GCLAMS	Dataset1	80.37 ± 2.24	90.70 ± 2.28	76.09 ± 3.42	67.52 ± 4.37	0.8688 ± 0.2143	0.8581 ± 0.2325
	Dataset2	87.36 ± 2.52	86.36 ± 0.66	69.29 ± 3.81	69.37 ± 2.58	0.9164 ± 0.3548	0.9027 ± 0.1429
ELDMA	Dataset1	87.63 ± 2.08	88.84 ± 1.47	86.01 ± 3.46	74.91 ± 4.12	0.9447 ± 0.0015	0.9444 ± 0.1236
	Dataset2	88.93 ± 0.74	91.31 ± 1.68	86.57 ± 1.93	77.99 ± 1.45	0.9558 ± 0.3484	0.9485 ± 0.7123
PSRR	Dataset1	79.59 ± 1.67	78.47 ± 2.32	80.7 ± 2.43	59.22 ± 3.33	0.8689 ± 0.0134	0.8583 ± 0.0125
	Dataset2	80.37 ± 1.54	78.32 ± 1.94	82.4 ± 3.63	60.84 ± 3.21	0.8765 ± 0.0135	0.8722 ± 0.0133
MHESMMR	Dataset1	90.55 ± 0.91	92.73 ± 1.65	88.38 ± 2.03	81.21 ± 1.80	0.9594 ± 0.0065	0.9611 ± 0.0046
	Dataset2	90.97 ± 1.51	92.74 ± 1.76	89.21 ± 3.25	82.05 ± 2.94	0.9620 ± 0.0099	0.9649 ± 0.0062

more intuitively, the result of the MHESMMR model and other models are compared in Figs. 10 and 11. According to Table 4, the AUC values of the MHESMMR model are 8.72% higher than the PSRR model in up-regulation pairs and 9.04% higher than the PSRR model in down-regulation pairs. The AUPR values of the MHESMMR model are 10.03% higher than the PSRR model in up-regulation pairs and 8.74% higher than the PSRR model in down-regulation pairs. These results clarified that the MHESMMR model, with the benefit from attributed multiplex heterogeneous network embedding,

could be an accurate and efficient computational model for the prediction of the regulation of miRNAs expression by SM on a large scale.

Case study

For validating the performance of MHESMMR model on predicting potentially the regulation of miRNAs expression by SM. We conducted a case study identifying miRNA targets of specific drugs. 5-FU (CID 3385) was selected as the designated drug for this case study. 5-FU is a kind of common chemotherapy drugs for cancer [50]. It can inhibit the proliferation of cancer cells by changing the metabolism of RNA and DNA to reduce the synthesis of specific proteins [51–54]. Therefore, we utilized the Dataset2 to construct the down-regulation pairs prediction model. We removed all of relation pairs between 5-FU and all of miRNAs in Dataset2 and then implement MHESMMR model based on rest SM-miRNA relation pairs. The prediction results are shown in Table 5. According to the Table 5, among the potential 5-fu-related miRNAs with the top 10 highest prediction scores, seven of them were proved by the PubMed literature to be inhibited by 5-FU.

For instance, Shah et al. demonstrated 5-FU can down-regulate the expression of hsa-miR-21-5p by qRT-PCR. Hsa-miR-30a-5p can medita the effect of 5-FU on p53-mutatant cells which is resistant to 5-FU. MiRNAs affected by 5-FU can target important p53 regulatory genes. Zhou et al. had identified hsa-miR-92a-3p, hsa-miR-15b-5p, hsa-miR-191-5p and hsa-miR-128-3p as down-regulate in HCT-8 and HCT-116 colon cancer cells after exposure to 5-FU by microarray analysis [55]. Rossi et al. discovered down-regulation of hsa-miR-210-3p in 5-FU treated HC.21 cell lines.

Conclusion

It is well known that the abnormal expression of miRNAs is an important role in various pathological processes. Through SM drugs, the oncomiRs can be down-regulated and the TSmiRs can be up-regulated. Therefore, an efficient miRNA drug regulatory relationship prediction model is needed. In this study, we developed an innovative computational method for the prediction of regulation relation between miRNA-SM based on graph embedding and machine learning named MHESMMR. It combines the LINE algorithm, the GATNE algorithm and the LightGBM method. And it shows the usefulness of non-linear relationships in identifying the potential miRNA-SM associations.

Table 5 The top 10 predicted miRNAs interacted with the 5-FU

Rank	PubChem ID	miRNA	Possibility	Evidence
1	3385	hsa-miR-21-5p	0.9857	21,506,117
2	3385	hsa-miR-92a-3p	0.9733	19,956,872
3	3385	hsa-miR-16-5p	0.9660	Unconfirmed
4	3385	hsa-miR-15b-5p	0.9641	19,956,872
5	3385	hsa-miR-128-3p	0.9555	19,956,872
6	3385	hsa-miR-30a-5p	0.9542	21,506,117
7	3385	hsa-miR-155-5p	0.9518	Unconfirmed
8	3385	hsa-miR-191-5p	0.9385	19,956,872
9	3385	hsa-miR-210-3p	0.9317	17,702,597
10	3385	hsa-miR-107	0.9307	Unconfirmed

For evaluating the performance of the proposed model, we divide the up-regulation pairs and down-regulation pairs in the SM2miR dataset into Dataset1 and Dataset2. And we performed tests on these two datasets under a fivefold cross-validation. The MHESMMR model yielded average accuracies of 79.59% and 80.37% for Dataset1 and Dataset2, respectively. In addition, we compare the proposed model with another existing model to verify the predictive ability of our model. We also compare the LightGBM method with other classical machine learning classifiers. The experimental results demonstrated that the MHESMMR model is a valuable tool to predict miRNA-SM regulation relations. In the future, we intend to search for more effective feature extraction methods and develop diverse feature descriptors to construct better prediction models.

Acknowledgements
The authors thank the anonymous reviewers for their valuable suggestions

Author contributions
Y.-J.G., Z.-H.Y., L.-P.L., and C.-Q.Y.: conceptualization, methodology, software, validation, resources and data curation. M.-M.W., and C. Y.: preparing the resource, performed bioinformatics and data analysis. X.-F.W., and L.-X.G.: writing original draft preparation. All authors contributed to manuscript revision. All authors have read and agreed to the published version of the manuscript.

Funding
This work was supported by National Natural Science Foundation of China under Grants 62273284. The authors would like to thank all the editors and anonymous reviewers for their work.

Availability of data and materials
SM2miR [SM2miR (jianglab.cn)] is freely available to the public without registration or login requirements. The data and source code can be found at https: //github.com/Heath0/ MHESMMR.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 7 September 2023 Accepted: 21 December 2023
Published online: 02 January 2024

References

1. Cai Y, Yu X, Hu S, et al. A brief review on the mechanisms of miRNA regulation. Genom Proteom Bioinf. 2009;7:147–54.
2. Lee R, Feinbaum R, Ambros V. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. Cell. 1993;75:843–54.
3. Chen L, Heikkinen L, Wang C, et al. Trends in the development of miRNA bioinformatics tools. Brief Bioinf. 2019;20:1836–52.
4. Meister G, Tuschl TJN. Mechanisms of gene silencing by double-stranded RNA. Nature. 2004;431:343–9.
5. Lindsay MA, Griffiths-Jones S, Dalmay TJE. Mechanism of miRNA-mediated repression of mRNA translation. Essays Biochem. 2013;54:29–38.
6. Weiss CN, Ito K. A macro view of microRNAs: the discovery of microRNAs and their role in hematopoiesis and hematologic disease. Int Rev Cell Mol Biol. 2017;334:99–175.
7. Bartel DPJC. Metazoan micrornas. Cell. 2018;173:20–51.
8. Bartel DPJ. MicroRNAs: genomics, biogenesis, mechanism, and function. Cell. 2004;116:281–97.
9. Baranwal S, Alahari SKJI. miRNA control of tumor cell invasion and metastasis. Int J Cancer. 2010;126:1283–90.
10. DeSano JT, Xu LJTAT. MicroRNA regulation of cancer stem cells and therapeutic implications. AAPS J. 2009;11:682–92.
11. Fu SW, Chen L, Man YJ. miRNA biomarkers in breast cancer detection and management. J Cancer. 2011;2:116.
12. Li J, Yang X, Guan H, et al. Exosome-derived microRNAs contribute to prostate cancer chemoresistance. Int J Oncol. 2016;49:838–46.
13. Rossi M, Amodio N,Teresa Di Martino M, et al. From target therapy to miRNA therapeutics of human multiple myeloma: theoretical and technological issues in the evolving scenario. Curr Drug Targets. 2013;14:1144–9.

14. Naro Y, Thomas M, Stephens MD, et al. Aryl amide small-molecule inhibitors of microRNA miR-21 function. Bioorg Med Chem Lett. 2015;25:4793–6.
15. Schmidt MF. Drug target miRNAs: chances and challenges. Trends Biotechnol. 2014;32:578–85.
16. Chandrasekhar S, Pushpavalli SN, Chatla S, et al. aza-Flavanones as potent cross-species microRNA inhibitors that arrest cell cycle. Bioorg Med Chem Lett. 2012;22:645–8.
17. Hopkins AL, Groom CR. The druggable genome. Nat Rev Drug Discov. 2002;1:727–30.
18. Overington JP, Al-Lazikani B, Hopkins AL. How many drug targets are there? Nat Rev Drug Discov. 2006;5:993–6.
19. Zhang S, Chen L, Jung EJ, et al. Targeting microRNAs with small molecules: from dream to reality. Clin Pharmacol Ther. 2010;87:754–8.
20. Disney MD, Winkelsas AM, Velagapudi SP, et al. Inforna 2.0: a platform for the sequence-based design of small molecules targeting structured RNAs. ACS Chem Biol. 2016;11:1720–8.
21. Bose D, Jayaraj GG, Kumar S, et al. A molecular-beacon-based screen for small molecule inhibitors of miRNA maturation. ACS Chem Biol. 2013;8:930–8.
22. Chen X, Guan N-N, Sun Y-Z, et al. MicroRNA-small molecule association identification: from experimental results to computational models. Brief Bioinf. 2020;21:47–61.
23. Lv Y, Wang S, Meng F, et al. Identifying novel associations between small molecules and miRNAs based on integrated molecular networks. Bioinformatics. 2015;31:3638–44.
24. Wang J, Meng F, Dai E, et al. Identification of associations between small molecule drugs and miRNAs based on functional similarity. Oncotarget. 2016;7:38658.
25. Liu F, Peng L, Tian G, et al. Identifying small molecule-miRNA associations based on credible negative sample selection and random walk. Front Bioeng Biotechnol. 2020;8:131.
26. Guan N-N, Sun Y-Z, Ming Z, et al. Prediction of potential small molecule-associated microRNAs using graphlet interaction. Front Pharmacol. 2018;9:1152.
27. Qu J, Chen X, Sun Y-Z, et al. Inferring potential small molecule–miRNA association based on triple layer heterogeneous network. J Cheminf. 2018;10:1–14.
28. Yu D-L, Yu Z-G, Han G-S, et al. Heterogeneous types of miRNA-disease associations stratified by multi-layer network embedding and prediction. Biomedicines. 2021;9:1152.
29. Liu X, Wang S, Meng F, et al. SM2miR: a database of the experimentally validated small molecules' effects on microRNA expression. Bioinformatics. 2013;29:409–11.
30. Rodriguez JJ, KunchevaAlonso LICJ, et al. Rotation forest: a new classifier ensemble method. IEEE Trans Pattern Anal Mach Intell. 2006;28:1619–30.
31. Cereto-Massagué A, Ojeda MJ, Valls C, et al. Molecular fingerprint similarity search in virtual screening. Methods. 2015;71:58–63.
32. Weininger D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. J Chem Inf Comput Sci. 1988;28:31–6.
33. Tang J, Qu M, Wang M et al. Line: large-scale information network embedding. In: Proceedings of the 24th international conference on world wide web. 2015, p. 1067–1077.
34. Gu W, Tandon A, Ahn Y-Y, et al. Principled approach to the selection of the embedding dimension of networks. Nat Commun. 2021;12:1–10.
35. Dong Y, Chawla NV, Swami A. metapath2vec: scalable representation learning for heterogeneous networks. In: Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining. 2017, p. 135–144.
36. Mikolov T, Chen K, Corrado G et al. Efficient estimation of word representations in vector space 2013.
37. Liu W, Lin H, Huang L et al. Identification of miRNA–disease associations via deep forest ensemble learning based on autoencoder. Brief Bioinf 2022;23:bbac104.
38. Liu W, Tang T, Lu X et al. MPCLCDA: predicting circRNA–disease associations by using automatically selected metapath and contrastive learning. Brief Bioinf 2023:bbad227.
39. Friedman J. Greedy function approximation: a gradient boosting machine. Ann Stat 2001:1189–1232.
40. Ke G, Meng Q, Finley T et al. Lightgbm: a highly efficient gradient boosting decision tree. Adv Neur Inf Process Syst 2017;30.
41. Metz CE. Basic principles of ROC analysis. In: Seminars in nuclear medicine. 1978, p. 283–298. Elsevier.
42. Bradley A. The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recognit. 1997;30:1145–59.
43. Guo Z-H, You Z-H, Huang D-S, et al. A learning based framework for diverse biomolecule relationship prediction in molecular association network. Commun Biol. 2020;3:1–9.
44. Cen Y, Zou X, Zhang J et al. Representation learning for attributed multiplex heterogeneous network. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2019, p. 1358–1368.
45. Pokuri S, Education M. A hybrid approach for feature selection analysis on the intrusion detection system using Navi Bayes and improved BAT algorithm. Turk J Comput Math Educ (TURCOMAT). 2021;12:5078–87.
46. LaValley MPJC. Logistic regression. Circulation. 2008;117:2395–9.
47. Hearst MA, Dumais ST, Osuna E, et al. Support vector machines. IEEE Intell Syst Appl. 1998;13:18–28.
48. Qi Y. Random forest for bioinformatics. Ensemble machine learning. Springer, 2012;307–323.
49. Yu F, Li B, Sun J, et al. PSRR: a web server for predicting the regulation of miRNAs expression by small molecules. Front Mol Biosci. 2022;9:817294.
50. Windle R, Bell P, Shaw D. Five year results of a randomized trial of adjuvant 5-fluorouracil and levamisole in colorectal cancer. J Br Surg. 1987;74:569–72.
51. Hernández-Vargas H, Ballestar E, Carmona-Saez P, et al. Transcriptional profiling of MCF7 breast cancer cells in response to 5-Fluorouracil: Relationship with cell cycle changes and apoptosis, and identification of novel targets of p53. Int J Cancer. 2006;119:1164–75.

52. Rossi L, Bonmassar E, Faraoni I. Modification of miR gene expression pattern in human colon cancer cells following exposure to 5-fluorouracil in vitro. Pharmacol Res. 2007;56:248–53.
53. Shah MY, Pan X, Fix LN, et al. 5-fluorouracil drug alters the microrna expression profiles in MCF-7 breast cancer cells. J Cell Physiol. 2011;226:1868–78.
54. Bash-Imam Z, Thérizols G, Vincent A, et al. Translational reprogramming of colorectal cancer cells induced by 5-fluorouracil through a miRNA-dependent mechanism. Oncotargets. 2017;8:46219.
55. Zhou J, Zhou Y, Yin B, et al. 5-Fluorouracil and oxaliplatin modify the expression profiles of microRNAs in human colon cancer cells in vitro. Oncol Rep. 2010;23:121–8.

Publisher's Note