Molecular Therapy
Nucleic Acids
Original Article

Joint masking and self-supervised strategies for inferring small molecule-miRNA associations

Zhecheng Zhou,[1] Linlin Zhuo,[1] Xiangzheng Fu,[2] Juan Lv,[3] Quan Zou,[4] and Ren Qi[5,6]

[1]School of Data Science and Artificial Intelligence, Wenzhou University of Technology, Wenzhou 325027, China; [2]College of Computer Science and Electronic Engineering, Hunan University, Changsha 410012, China; [3]College of Traditional Chinese Medicine, Changsha Medical University, Changsha 410000, China; [4]Institute of Fundamental and Frontier Sciences, University of Electronic Science and Technology of China, Chengdu 611730, China; [5]Yangtze Delta Region Institute (Quzhou), University of Electronic Science and Technology of China, Quzhou, Zhejiang, China; [6]School of Life Science and Technology, University of Electronic Science and Technology of China, Chengdu, China

Inferring small molecule-miRNA associations (MMAs) is crucial for revealing the intricacies of biological processes and disease mechanisms. Deep learning, renowned for its exceptional speed and accuracy, is extensively used for predicting MMAs. However, given their heavy reliance on data, inaccuracies during data collection can make these methods susceptible to noise interference. To address this challenge, we introduce the joint masking and self-supervised (JMSS)-MMA model. This model synergizes graph autoencoders with a probability distribution-based masking strategy, effectively countering the impact of noisy data and enabling precise predictions of unknown MMAs. Operating in a self-supervised manner, it deeply encodes the relationship data of small molecules and miRNA through the graph autoencoder, delving into its latent information. Our masking strategy has successfully reduced data noise, enhancing prediction accuracy. To our knowledge, this is the pioneering integration of a masking strategy with graph autoencoders for MMA prediction. Furthermore, the JMSS-MMA model incorporates a node-degree–based decoder, deepening the understanding of the network's structure. Experiments on two mainstream datasets confirm the model's efficiency and precision, and ablation studies further attest to its robustness. We firmly believe that this model will revolutionize drug development, personalized medicine, and biomedical research.

INTRODUCTION

MicroRNA (miRNA) represents a category of noncoding RNA molecules,[1] typically composed of approximately 20–25 nt. These molecules play a pivotal role in the regulation of gene expression.[2] Small molecules refer to some relatively small organic molecules, including compounds and metabolites.[3] A wealth of research has demonstrated the critical role that miRNA plays in numerous aspects of human life, such as gene expression regulation,[4] cell-cycle control,[5] development and organogenesis,[6] immune response regulation,[7] metabolic control,[8] and even the initiation and progression of tumors.[9] It is precisely because of the key role of miRNA in the human body that more and more researchers are devoting themselves to the invention of miRNA-targeted drugs.[10] However, the traditional experimental method is complicated and requires a lot of time and labor costs.[11] Therefore, it is imminent to develop an efficient model to analyze and calculate the possible association between small molecules and miRNA.

Traditional miRNA analysis methods mainly include miRNA sequence analysis,[12,13] which uses high-throughput sequencing technology to comprehensively detect and quantify miRNA in cells or tissues. This method can provide a comprehensive miRNA expression profile, compare miRNA differences between different samples, and effectively analyze miRNA. Real-time qPCR,[14] which is a commonly used quantitative method for miRNA expression, is based on the principle of DNA synthesis and amplification, can quickly and highly sensitively measure the expression level of specific miRNA, and can quantitatively analyze the expression difference of miRNA under different conditions. Northern blot,[15] which is a traditional miRNA detection method, is used to analyze the size and expression of miRNA. It separates the total RNA by electrophoresis, transfers it to the membrane, and then uses the labeled miRNA probe for hybridization detection to determine the presence and relative expression level of the target miRNA. Although these traditional methods are reliable, they are costly in manpower and time.

Recently, with the vigorous development of computer technology and the emergence of machine learning–based methods,[16] researchers have developed related models to solve different biological problems,[17,18] such as predicting noncoding RNA (ncRNA)–protein interactions (NPIs), miRNA-disease association (MDAs), and so on. For

Received 26 September 2023; accepted 13 December 2023;
https://doi.org/10.1016/j.omtn.2023.102103.

Correspondence: Linlin Zhuo, School of Data Science and Artificial Intelligence, Wenzhou University of Technology, Wenzhou 325027, China.
E-mail: 20210339@wzut.edu.cn
Correspondence: Xiangzheng Fu, College of Computer Science and Electronic Engineering, Hunan University, Changsha 410012, China.
E-mail: fxz326@hnu.edu.cn
Correspondence: Ren Qi, Yangtze Delta Region Institute (Quzhou), University of Electronic Science and Technology of China, Quzhou, Zhejiang, China.
E-mail: qiren@tju.edu.cn

instance, Zhou et al. used a deep multihead attention mechanism to mine the information of ncRNA and protein, so as to accurately predict the NPIs.[19] Liu et al. adopted a deep autoencoder-based forest ensemble learning strategy to effectively predict the MDAs.[20] Wei et al. combined contrastive learning with graph neural networks to enrich the feature representations of drugs and foods to accurately predict their interactions.[21] These methods have achieved excellent performance on multiple association prediction tasks, but they cannot be directly used for small molecule-miRNA associations MMA prediction tasks. This is mainly due to the large differences in data sources and associated networks. These models need to be appropriately adjusted before they can be used to infer potential MMAs. Inspired by these works, the MMA prediction model was also promoted.

Many methods based on matrix completion and machine learning to predict the MMAs have been developed. Wang et al. proposed a random forest–based model and integrated multiple similarity features to predict the MMAs.[22] Luo et al. proposed a non-negative matrix decomposition model to discover the unknown MMAs.[23] On the basis of this, Ni et al. combined layer attention network and matrix decomposition to predict the MMAs.[24] Peng et al. proposed a model based on deep autoencoder and extensible boosting tree to predict the MMAs, making contributions to biological research.[25] Wang et al. adopted matrix decomposition to calculate the potential representations of small molecules and miRNAs and then calculate the inner product to score the MMA pairs. The main highlight is that the missing values of the incidence matrix can be preprocessed using the WKNKN method.[26] Wang et al., on the basis of the truncated Schatten p-norm, developed an MMA prediction model, and experimental results demonstrate its advanced performance.[27] These methods speed up MMA predictions, but they rely on artificially constructed features, making their performance less robust.

Overall, deep learning models performed well in identifying the potential MMAs. However, the performance of these models often suffers from noisy data, insufficient feature extraction, and so forth. The self-supervised learning strategy can automatically extract supervision signals from input data to perform self-training. Inspired by this, we proposed an MMA prediction model based on the graph auto encoder (GAE) framework. The model follows the rules of self-supervised learning and simultaneously absorbs information from small molecules and miRNA itself, as well as topological information from the MMA network. Then, the proposed model used an edge decoder and a node decoder to reconstruct the graph, which can obtain richer and more robust representations from unlabeled graph data. At the same time, we designed a masking strategy based on Bernoulli distribution to mask some edges for self-supervised training when modeling graph data, which effectively mitigate the impact of noise data. Overall, our contributions can be summarized as the following:

(1) We developed a graph masked autoencoder–based MMA prediction model that can predict unknown MMAs quickly and accurately.

(2) We designed a masking strategy for graph autoencoder training, which effectively improves training efficiency and mitigates the impact of noisy data.

(3) We designed a degree-based node decoder that can efficiently learn the latent structure of the small molecule-miRNA graph.

(4) We constructed multiple sets of experiments to verify the performance of the model, and verified the role of each part of the model through parameter experiments.

RESULTS

Experimental setup

We identified known MMAs in the datasets (as shown in Table 1) as positive samples and randomly picked the same number of negative samples. The datasets are divided into training set, verification set, and test set according to the ratio of 8:1:1. To prevent data leakage, the training set, validation set, and test set are independent and there are no common MMAs. That is, the MMAs in the training set will not appear in the validation set and test set. In the experiment, we set the input feature size of dataset1 to 990, the input feature size of dataset2 to 1,050, the hyperparameter α to 0.06, and the masking rate p to 0.4. In addition, the encoder was set to a two-layer graph isomorphism network (GIN), and the embedding sizes were set to 64 and 128, respectively; the decoder and edge decoder are both two-layer multilayer perceptrons (MLPs), and the output sizes are 128 and 64, respectively.

Evaluation indicators

In the experiments, we not only selected area under the receiver operating characteristic curve (AUC) and area under the precision-recall curve (AUPR) indicators but we also selected multiple indicators to comprehensively evaluate the performance of the model, including accuracy (Acc), sensitivity (Sen), precision (Pre), Spe, F1-score, and Matthews correlation coefficient (MCC). The calculation formulas for these indicators are as follows:

$$Acc = \frac{TP+TN}{TP+TN+FP+FN}, Sen = \frac{TP}{TP+FN},$$

$$Spe = \frac{TN}{TN+FP}, Pre = \frac{TP}{TP+FP},$$

$$F1 - score = 2 \cdot \frac{Pre \cdot Sen}{Pre+Sen},$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP) \times (TP+FN) \times (TN+FP) \times (TN+FN)}} \tag{Equation 1}$$

In the above equation, *TP* represents the number of positive MMAs correctly classified, *FP* represents the number of misclassified positive MMAs, *TN* represents the number of negative MMAs correctly classified, and *FN* represents the number of misclassified negative MMAs.

Performance evaluation

To verify the performance of our proposed model, we conducted 5- and 10-fold cross-validation experiments on two datasets. The dataset is randomly divided into 5 and 10 parts, one part is left as the test set for each experiment, and the remaining part is used as the training set.

Table 1. Statistical information of the datasets

Datasets	Small molecules	miRNAs	Associations
dataset1	831	541	664
dataset2	39	286	664

This strategy can eliminate the impact of the randomness of the dataset division on the experiment. The AUC curve of the model is shown in Figures S1 and S2. In the 5-fold cross-validation experiment, the proposed model obtained an average AUC performance of 99.16% on dataset1 and an average AUC performance of 96.26% on dataset2. In the 10-fold cross-validation experiment, the proposed model obtained an average AUC performance of 99.90% on dataset1 and an average AUC performance of 97.10% on dataset2.

Similarly, we calculated other indicators in the 5-fold crossover experiment, and the results of each round are given in Table 2, and the average value is calculated. In addition, we selected DAESTB, which ranks second in AUC indicator, for comparative analysis. Under the same dataset division, the Acc, Pre, Sen, MCC, and F1-score indicators of the DAESTB[25] model were counted, and the results are shown in Table 2. Obviously, except for the Acc indicator, the Pre, Sen, MCC, and F1-score indicators of the DAESTB[25] model are much lower than these indicators of the proposed JMSS-MMA model. On the one hand, it may be because the DAESTB model focuses too much on the Acc indicator and ignores other indicators when adjusting parameters. On the other hand, it may be because the DAESTB model does not consider the imbalance of positive and negative samples.

Comparison with other models

We conducted comparative experiments with DAESTB,[25] EKRRSMMA,[28] GISMMA,[29] and BNNRSMMA.[30] The following is a brief introduction to these compared models:

(1) DAESTB[25]: This model is based on autoencoders and gradient boosted trees to identify potential MMAs. The similarity matrix of small molecules and miRNAs, as well as their association matrix data, are collected and input to an autoencoder for encoding. Then, the model uses gradient boosting trees to classify small molecule-miRNA pairs.

(2) EKRRSMMA[28]: This model is based on the method of ridge regression, combining dimensionality reduction techniques and ensemble learning to mine potential MMAs.

(3) GISMMA[29]: This model is based on the graphlet method, uses 28 isomers to describe the relationship between two small molecules (or miRNAs), and calculates the number of subgraphs interactions between the the small molecule similarity matrix and the miRNA similarity matrix to identify MMAs.

(4) BNNRSMMA[30]: This model identifies unknown MMAs based on bounded kernel norm regularization. The model integrates the similarity matrix of small molecules and miRNAs, and constructs a heterogeneous small molecule-miRNA association network. Then, it predicts whether there is an association between small molecules and miRNAs by minimizing its nuclear norm.

Our model is compared with DAESTB,[25] EKRRSMMA,[28] GISMMA,[29] and BNNRSMMA[30] models in dataset1 and dataset2, and the results are shown in Figure S3. Twenty experiments were carried out for each model to reduce the influence of random factors. The results in Figure S3 show that the JMSS-MMA and DAESTB models based on autoencoders achieve the better performance, and the other models achieve slightly worse performance. This shows that the autoencoder can fully extract the features of the small molecule-miRNA graph. Furthermore, our model achieves the best performance with the largest average AUC value. At the same time, the variance of the model performance is minimal. This shows that adding a mask strategy and a node decoder can make the model more robust and stable, which can alleviate the influence of noisy data in the graph and learn the underlying graph structure.

Table 2. Results of the 5-fold crossover experiment of the JMSS-MMA model on 2 datasets

Datasets	Testing set	Acc, %	Pre, %	Sen, %	Spe, %	F1-score, %	MCC, %
dataset1	1	97.7272	98.4615	96.9696	98.4848	97.7099	95.4655
	2	98.1061	99.2248	96.9697	99.2424	98.0843	96.2370
	3	98.8636	98.4962	99.2424	98.4848	98.8679	97.7300
	4	95.8333	96.1832	95.4545	96.2121	95.8175	91.6693
	5	98.8636	98.4962	99.2424	98.4848	98.8679	97.7301
	Average	97.8787	98.1723	97.5757	98.1817	97.8695	95.7663
	DAESTB	99.8732	68.5834	19.6942		30.5743	36.6864
dataset2	1	93.9394	89.7260	99.2424	88.6364	94.2446	88.3773
	2	92.8030	87.4172	99.2424	85.6061	93.2862	86.5069
	3	89.7727	84.3137	97.7273	81.8182	90.5263	80.5716
	4	92.0455	88.2759	96.9697	87.1212	92.4188	84.5017
	5	90.1515	83.5443	99.2424	80.3030	91.0345	81.9076
	Average	91.7424	86.6554	98.4848	84.6969	92.3020	84.3730
	DAESTB	94.4876	53.8464	16.6764		25.4554	27.8556

Table 3. Results of the ablation experiments of the JMSS-MMA model on 2 datasets

Datasets	Models	AUC	AUPR	Acc	Pre	F1-score	MCC
dataset1	w/o mask	80.1768%	89.2827%	88.2576%	99.2424%	86.6953%	78.7167%
	w/o GAE	77.1768%	85.2827%	84.2576%	94.1176%	82.6953%	72.7137%
	JMSS-MMA	99.8508%	99.8719%	99.2424%	99.2424%	99.2424%	98.4848%
dataset2	w/o mask	91.8905%	90.7293%	72.7273%	94.1176%	64.0000%	51.9719%
	w/o GAE	88.8905%	87.7293%	69.5142%	91.5132%	62.5124%	49.4233%
	JMSS-MMA	97.0156%	95.9959%	93.1818%	91.3043%	93.3333%	86.4530%

In addition to AUC, we further compared various indicators with the DAESTB model and conducted experiments on both datasets. In the experiment, we divided the positive and negative proportions of the samples input into the model to be the same as the DAESTB setting. We used the heatmap to represent the final experimental results in Figure S4. From the experimental results, our model is superior to the DAESTB model in all indicators except Acc, which further illustrates the superiority of our model.

Ablation experiments

We conducted experiments to explore the impact of GAE technology and masking strategy on model performance. In the experiments, we maintained the single variable principle. Table 3 presents the results of the ablation experiments. In Table 3, ‘w/o mask’ means that the model does not perform masking operations, and ‘w/o GAE’ means that the model does not use GAE technology and only uses a graph convolutional network (GCN) as the encoding layer. It can be seen in Table 3 that without performing masking operations or using GAE technology, the model performance drops significantly. This shows that the use of GAE technology can enable the model to absorb the topological information of the MMA network and the feature of the node itself, and extract a robust node representation. In addition, performing partial masking operations on the MMA network can alleviate the impact of redundant data and improve model performance.

Parameter experiments

To explore the impact of parameters on model performance, we constructed multiple sets of parameter experiments.

Influence of α

For the node decoder, the hyperparameter α is used to adjust the weight of its loss. To explore the role of this decoder, we constructed experiments on the hyperparameter α. We changed the value of α on the two datasets to show the performance of the model. In the experiment, except for the variables to be verified, other parameters remained consistent. For example, the masking rate p is set to 0.4. The output sizes of the graph neural network (GNN) encoder are 64 and 128, respectively, and the output sizes of the decoder are 128 and 64, respectively. In dataset1, the initial feature dimension of the node is 990; in dataset2, the initial feature dimension of the node is 1,050.

The results are shown in Figure S5. In particular, $\alpha = 0$ means not to use the degree-based node decoder. At this point, the model achieves the worst performance. The results in Figure S5 prove that adding a decoder based on node degree can effectively improve the performance of the model, but the weight should not be too large. The value of the hyperparameter α is set between 0.002 and 0.01, and the model can obtain satisfactory performance. When setting the value of the hyperparameter α to 1, the effect drops significantly. Therefore, designing a degree-based node decoder can help improve model performance. Moreover, the range of hyperparameter α that can be selected by the model is relatively large, and the determination of the parameters is very convenient.

Influence of mask ratio

We designed a masking strategy based on the Bernoulli distribution to partially mask the input small molecule-miRNA association. To verify the effect of the masking strategy, we designed parametric experiments on the masking ratio. Keeping other conditions unchanged, different shading ratios were set to test the performance of the model. In the experiment, α is set to 0.01, the output sizes of the GNN encoder are 64 and 128, and the output sizes of the decoder are 128 and 64. In dataset1, the initial feature dimension of the node is 990; in dataset2, the initial feature dimension of the node is 1,050. The AUC and AUPR were used as evaluation indicators, and the model without masking strategy was used as the baseline method.

The experimental results in Figure S6 indicate that the performance of the models with masking are significantly superior to that of the model without masking. This proves that the proposed masking strategy can effectively improve the performance of the model, probably because of slowing down the impact of noise in the graph data. In dataset1, the model achieved better performance when the occlusion rate was 0.1–0.8; in dataset2, the model achieved better performance when the occlusion rate was 0.3–0.8. Different masking ratios can affect the performance of the model, but the range of parameters that can be selected is large. This makes it easy to determine suitable parameters.

Influence of the GNN layer

In the proposed JMSS-MMA model, the graph encoder is optional and can be determined as various GNN models. To study the impact of different GNN encoders on model performance, we selected three GNN models (GIN, GCN, and sample and aggregate [SAGE]),[31–33] and conducted experiments on dataset1 and dataset2, respectively. In Table 4, it can be seen from the results that in dataset1, the model

Table 4. Results of different GNN encoders of the JMSS-MMA model on 2 datasets

Datasets	GNN encoders	AUC	AUPR	Acc	Pre	F1-score	MCC
dataset1	GIN	99.8508%	99.8719%	99.2424%	99.2424%	99.2424%	98.4848%
	GCN	99.7188%	99.7705%	98.8636%	98.4962%	98.8679%	97.7301%
	GraphSAGE	99.9311%	99.9292%	99.2424%	99.2424%	99.2424%	98.4848%
dataset2	GIN	97.0156%	95.9959%	93.1818%	91.3043%	93.3333%	86.4530%
	GCN	97.2854%	96.2432%	93.1818%	90.7143%	93.3824%	86.5227%
	GraphSAGE	95.0413%	92.5449%	92.0455%	88.2759%	92.4188%	84.5017%

achieved the best performance using SAGE as the encoder and achieved suboptimal results using GIN. In dataset2, the model achieved the best performance using GCN and suboptimal results using GIN. In general, however, no matter which GNN encoder is used, the difference in model performance is small. This indicates that the proposed model can be adapted to different GNN encoders, so the appropriate GNN encoder can be selected freely and easily.

Case studies

We conducted case analysis experiments to further demonstrate the practical significance of the model. For a specific small molecule, the miRNA predicted to be associated with it, or for a specific miRNA, the small molecule predicted to be associated with it. We selected small molecules numbered CID: 3121 and CID: 5073, and miRNAs numbered hsa-mir-192 and hsa-mir-506 for verification.

The miRNA hsa-mir-192 is widely expressed in multiple tissues and cell types, including liver, kidney, lung, and stomach. It can affect key processes such as cell proliferation, apoptosis, differentiation, and metabolism by targeting and regulating the transcription and translation of multiple genes. In addition, it is also closely related to the occurrence and progression of various diseases such as tumor development, liver fibrosis, and cardiovascular disease. The miRNA hsa-mir-506 is involved in the regulation of various biological processes and diseases, and its expression level may be regulated by many factors, including physiological state, disease progression, and environmental stimuli. It affects key processes such as cell proliferation, apoptosis, differentiation, invasion, and metastasis by targeting and regulating the transcription and translation processes of multiple genes. In addition, it is closely related to the occurrence and progression of various diseases such as tumor development, cancer treatment resistance, and pulmonary fibrosis.

The study of these miRNAs provides insight into the pathogenesis of diseases and also provides new ideas for the early diagnosis and treatment of related diseases. In our experiment, we first removed these small molecules and miRNAs from the dataset, and then added them to the test set to test the prediction effect after training the model. Since the model output is a probability value, we sorted them. For small molecules, we selected the top 20 miRNAs that were predicted to be associated with them; for miRNAs, we selected the top 10 small molecules that were associated with them and compared these results with the real results in the database. The final results are shown in Tables 5, 6 and 7.

It can be seen from the results that after the prediction of our model, most of the associations have been verified in the database, which further proves the reliability of our model. For the small molecule numbered CID: 3121, most of the miRNAs predicted to be associated with it have been verified in the database. The only remaining miRNAs numbered hsa-mir-328, hsa-mir-181c, and hsa-mir-374a have not been verified in the database. However, there are relevant references[34–36] proving that the small molecule numbered CID: 3121 is indeed associated with these miRNAs. For the small molecule numbered CID: 5073, there is also a relevant reference[37] proving its association with hsa-mir-135a. Although these miRNAs have not been verified in the database, for the miRNA numbered hsa-mir-192, it was also proven that there is an association with the small molecules numbered CID: 6013 and CID: 60953 predicted by the model.[38,39] Similar results were found for the miRNA numbered hsa-mir-506. Therefore, the proposed model can effectively infer potential MMAs and is expected to provide guidance for diagnosing diseases and developing treatment options.

DISCUSSION

Accurate identification of MMAs plays an important role in biological processes such as gene regulation and disease development, and mining more potential associations provide a better understanding of complex regulatory networks and signal transmission mechanisms such as gene-disease. In this study, we investigated several deep learning–based MMA recognition models. These models can show

Table 5. Top 20 CID: 3121-related miRNAs predicted by JMSS-MMA in dataset2

miRNAs	dataset2	miRNAs	dataset2
hsa-let-7a-1	definited	hsa-mir-124-1	definited
hsa-let-7a-2	definited	hsa-mir-124-2	definited
hsa-let-7a-3	definited	hsa-mir-124-3	definited
hsa-let-7b	definited	hsa-mir-18a	definited
hsa-let-7i	definited	hsa-mir-328	undefinited
hsa-mir-101-1	definited	hsa-mir-125a	definited
hsa-mir-101-2	definited	hsa-mir-181c	undefinited
hsa-mir-122	definited	hsa-mir-21	definited
hsa-mir-125b-1	definited	hsa-mir-31	definited
hsa-mir-125b-2	definited	hsa-mir-374a	undefinited

Table 6. Top 20 CID: 5073-related miRNAs predicted by JMSS-MMA in dataset2

miRNA	dataset2	miRNA	dataset2
hsa-mir-30c-1	definited	hsa-mir-660	definited
hsa-mir-337	definited	hsa-mir-744	definited
hsa-mir-34a	definited	hsa-mir-760	definited
hsa-mir-345	definited	hsa-mir-769	definited
hsa-mir-376a-1	definited	hsa-mir-9-1	definited
hsa-mir-376a-2	definited	hsa-mir-9-2	definited
hsa-mir-379	definited	hsa-mir-9-3	definited
hsa-mir-381	definited	hsa-mir-302b	undefinited
hsa-mir-382	definited	hsa-mir-135a	undefinited
hsa-mir-383	definited	hsa-mir-550a-1	undefinited

satisfactory results, but they rarely take into account the influence of noisy data, and there may be insufficient feature extraction. To this end, we propose an MMA identification model based on a graph autoencoder, which can more fully extract features by modeling the association graph of small molecules and miRNAs. We designed a masking strategy based on Bernoulli distribution, which can effectively remove noise in graph data and improve training efficiency. Meanwhile, we designed a degree-based node decoder that can effectively reveal the underlying graph structure. Taken together, the proposed model can accurately and efficiently identify potential MMAs. Multiple sets of experiments were constructed on public datasets, and the results verified the superior performance of the proposed model.

The proposed model can assist researchers in discovering more potential disease markers and targets, thereby deepening the understanding of disease mechanisms and providing new ideas and strategies for early diagnosis and treatment of diseases. The interaction between small molecules and miRNAs is also one of the important concerns in the drug development process. A predictive model is proposed to accurately predict the interaction between target small molecules and miRNA, which is expected to provide guidance for drug design and screening and accelerate the drug development process.

Table 7. Top 10 hsa-mir-192 and hsa-mir-506 small molecules predicted by JMSS-MMA in dataset2

small molecules (hsa-mir-192 related)	dataset2	small molecules (hsa-mir-506 related)	dataset2
CID:5757	definited	CID:5743	definited
CID:4095	definited	CID:446220	undefinited
CID:60823	definited	CID:31401	definited
CID:5311	definited	CID:36462	definited
CID:31703	definited	CID:5790	undefinited
CID:6013	undefinited	CID:31101	undefinited
CID:36462	definited	CID:60953	definited
CID:60700	definited	CID:4212	undefinited
CID:446220	definited	CID:60838	undefinited
CID:60953	undefinited	CID:448537	undefinited

MATERIALS AND METHODS

Materials

To validate our proposed model, we conducted experiments on two publicly available datasets. Table 1 shows the statistical information regarding these datasets.

Methods

To identify potential MMAs accurately and efficiently, we propose the JMSS-MMA model based on the graph autoencoder and masking strategy of the Bernoulli distribution. The model mainly includes two modules of data acquisition and model architecture, as shown in Figure 1. The first module obtains the similarity matrix of small molecules and miRNAs, which are described in detail in the Dataset section. The second module feeds these two similarity matrices into the graph autoencoder to learn the latent structure of the graph and finally identify unknown MMAs. The following describes the model architecture module in detail.

Model architecture

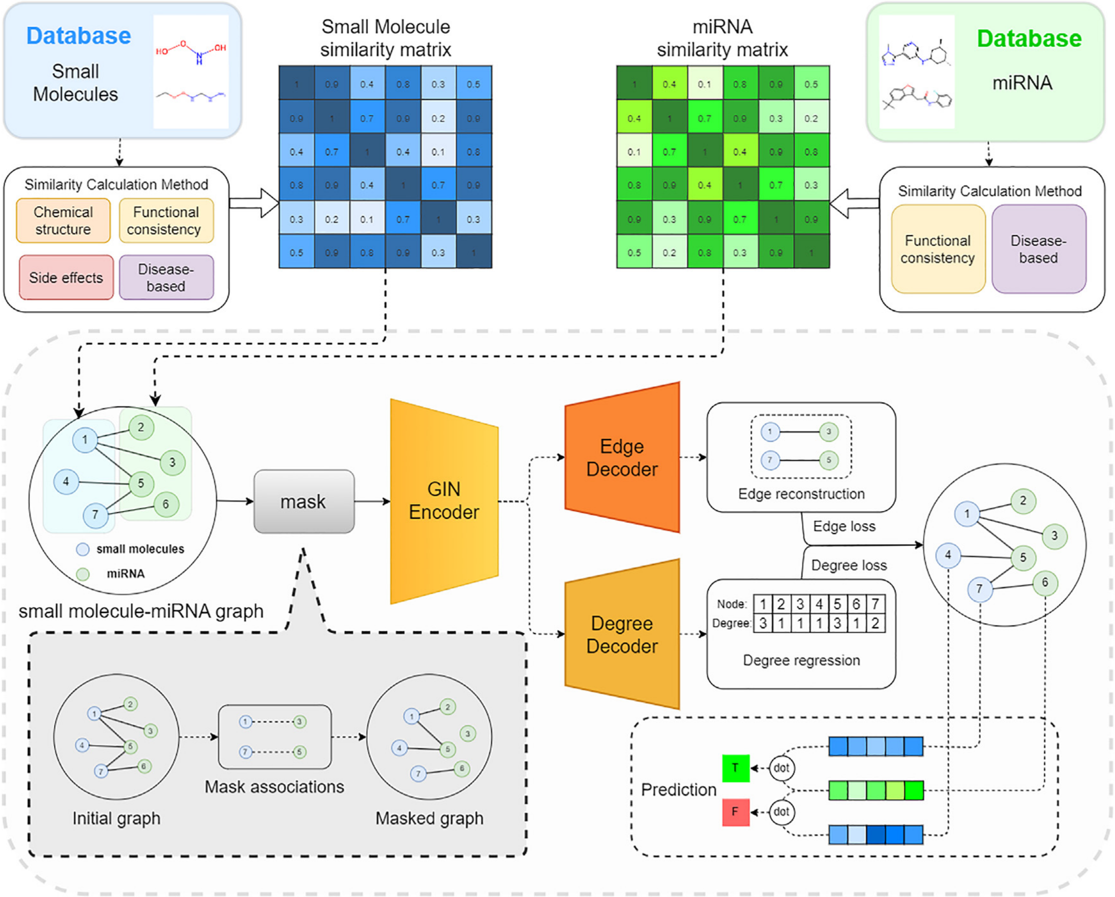
The JMSS-MMA model mainly adopts the basic architecture of the graph autoencoder.[40] Graph autoencoders are well suited for small molecule-miRNA interaction graphs and can preserve and learn the structural information of graphs. Our model is similar to traditional autoencoders, mainly consisting of two parts: encoder and decoder. In addition, we designed a masking strategy based on Bernoulli distribution, which can effectively alleviate the influence of noisy data in the graph. Also, we designed a degree-based decoder that can better learn the latent structure of graphs.

GNN encoder

In our model, GIN is adopted as the encoder (GCN and GraphSAGE are optional), and its core idea is to update the node representation by iteratively aggregating the node's neighbor information. In the task of modeling the small molecule-miRNA relationship, each small molecule node (or miRNA node) is aggregated by its own features and the features of the associated miRNA nodes at each iteration. MLP is adopted to update the aggregated node representations. Finally, the representation of each node is pooled to obtain a representation of the entire small molecule-miRNA graph. The iterative process of small molecule node features can be expressed as:

$$S_a^k = MLP^k\left(\left(1+\varepsilon^k\right)\cdot S_a^{k-1} + \sum_{b\in N(a)} M_b^{k-1}\right), \qquad \text{(Equation 2)}$$

where S_a^k denotes the feature representation of the small molecule at the k-th iteration, and MLP^k denotes the MLP operation applied at the k-th iteration. ε^k is a learnable scalar parameter used to balance the contribution of its own features and neighbor features. $N(a)$ represents the neighbors of the small molecule node a. M_b^{k-1} represents the feature representation of the miRNA b after the k-1-th iteration. The update process of miRNA node features is similar, and the calculation is as follows:



1	0.9	0.4	0.8	0.3	0.5
0.9	1	0.7	0.9	0.2	0.9
0.4	0.7	1	0.4	0.1	0.8
0.8	0.9	0.4	1	0.7	0.9
0.3	0.2	0.1	0.7	1	0.3
0.5	0.9	0.8	0.9	0.3	1

1	0.4	0.1	0.8	0.9	0.5
0.4	1	0.7	0.9	0.3	0.2
0.1	0.7	1	0.4	0.9	0.8
0.8	0.9	0.4	1	0.7	0.3
0.9	0.3	0.9	0.7	1	0.9
0.5	0.2	0.8	0.3	0.9	1

Node:	1	2	3	4	5	6	7
Degree:	3	1	1	1	3	1	2

Figure 1. Flowchart of the JMSS-MMA model

$$M_b^k = MLP^k\left(\left(1+\varepsilon^k\right)\cdot M_b^{k-1}+\sum_{a\in N(b)} S_a^{k-1}\right). \quad \text{(Equation 3)}$$

Decoder

The decoder mainly consists of two parts: one is an edge decoder and the other is a node encoder. They both consist of two layers of MLPs. The role of the edge decoder is to reconstruct the adjacency matrix of the graph—that is, to reconstruct the small molecule-miRNA association matrix. Specifically, after concatenating the embeddings of small molecule-miRNA pairs, their scores are calculated by MLP to predict whether the association exists. During the training process, the reconstructed MMAs and the real MMAs are used to calculate the loss value with the BCE function:

$$Edge_{loss} = (y-1)\cdot\log(1-p) - y\cdot\log(p), \quad \text{(Equation 4)}$$

where p represents the predicted value, that is, the probability of the existence of the MMA, and the value is between 0 and 1; y represents the real value (i.e., whether there is this edge, represented by 0 or 1). The function of the node encoder is to reconstruct the degree of each node (i.e., infer the number of miRNA nodes, or small molecule nodes, associated with the small molecule node, or miRNA node). During the training process, the predicted node degree and the real node degree are used to calculate the loss value with the mean squared error loss function:

$$Degree_{loss} = \frac{1}{x}\sum_{i=1}^{x}\left(y_i - p_i\right)^2, \quad \text{(Equation 5)}$$

where x represents the number of all nodes, y_i represents the actual degree of the i-th node, and p_i represents the predicted degree of the node. Adding the loss values of the two linearly, we obtain the final loss value representation of the model:

$$\varsigma_{total} = Edge_{loss} + \alpha Degree_{loss}, \quad \text{(Equation 6)}$$

where α is an adjustable hyperparameter used to balance the weights between the two losses.

Masked strategy

To mitigate the effect of noise in the small molecule-miRNA graph, we devised a probability distribution-based masking strategy. The core idea is to mask some associations in the small molecule-miRNA graph following a probability distribution. In each round of training, the model will follow the Bernoulli distribution and mask some MMAs on the small molecule-miRNA graph. Subsequently, edge decoders and degree decoders work together to reconstruct these masked MMAs. This strategy can effectively remove some of the noise present in the data. Specifically, we sample the set of known associations according to the Bernoulli distribution:

$$\zeta_{mask} \sim Bernoulli(p), \qquad \text{(Equation 7)}$$

where p is a probability value between 0 and 1, indicating the masked ratio of the graph, and different numbers of edges are masked by setting p values of different sizes.

Prediction

After the above process, the model reconstructs the small molecule-miRNA graph. We can obtain the representation of small molecule nodes and miRNA nodes, and use the dot product to calculate the probability of an association between them:

$$P_{i,j} = S_i^T M_j, \qquad \text{(Equation 8)}$$

where S_i represents the final representation of the i-th small molecule, M_j represents the final representation of the j-th miRNA, and $P_{i,j}$ represents the probability of an association between them.

DATA AND CODE AVAILABILITY

Our code and data are publicly available in the GitHub repository: https://github.com/ZZCrazy00/JMSSMMA.

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.omtn.2023.102103.

ACKNOWLEDGMENTS

The work was supported in part by the National Natural Science Foundation of China (nos. 62302339, 62002111, and 62372158), the Natural Science Foundation of Hunan Province (no. 2022JJ40090), and it was also supported by a grant from the Macao Young Scholars Program (project code: AM2021025).

AUTHOR CONTRIBUTIONS

Z.Z. and L.Z. were responsible for the experimental design and manuscript writing. X.F., J.L., R.Q., and Q.Z. were responsible for the corresponding guidance or revision.

DECLARATION OF INTERESTS

The authors declare no competing interests.

REFERENCES

1. Neilson, J.R., Zheng, G.X., Burge, C.B., and Sharp, P.A. (2007). Dynamic regulation of miRNA expression in ordered stages of cellular development. Genes Dev. *21*, 578–589.
2. Cai, Y., Yu, X., Hu, S., and Yu, J. (2009). A brief review on the mechanisms of mirna regulation. Dev. Reprod. Biol. *7*, 147–154.
3. Dervan, P. (2001). Molecular recognition of DNA by small molecules. Bioorg. Med. Chem. *9*, 2215–2235.
4. Gottesfeld, J.M., Neely, L., Trauger, J.W., Baird, E.E., and Dervan, P.B. (1997). Regulation of gene expression by small molecules. Nature *387*, 202–205.
5. Crews, C.M., and Mohan, R. (2000). Small-molecule inhibitors of the cell cycle. Curr. Opin. Chem. Biol. *4*, 47–53.
6. Asli, N.S., Pitulescu, M.E., and Kessel, M. (2008). MicroRNAs in organogenesis and disease. Curr. Mol. Med. *8*, 698–710.
7. Tsitsiou, E., and Lindsay, M.A. (2009). Micrornas and the immune response. Curr. Opin. Pharmacol. *9*, 514–520.
8. Krützfeldt, J., and Stoffel, M. (2006). MicroRNAs: A new class of regulatory genes affecting metabolism. Cell Metabol. *4*, 9–12.
9. Zhu, J., Zheng, Z., Wang, J., Sun, J., Wang, P., Cheng, X., Fu, L., Zhang, L., Wang, Z., and Li, Z. (2014). Different miRNA expression profiles between human breast cancer tumors and serum. Front. Genet. *5*, 149.
10. Miroshnichenko, S., and Patutina, O. (2019). Enhanced inhibition of tumorigenesis using combinations of mirna-targeted therapeutics. Front. Pharmacol. *10*, 488.
11. Ahmad, A., Ginnebaugh, K.R., Li, Y., Bao, B., Gadgeel, S.M., and Sarkar, F.H. (2014). Mirna targeted therapy in lung cancer. MicroRNA Target. Cancer Ther. *2014*, 99–114.
12. Aldridge, S., and Hadfield, J. (2012). Introduction to mirna profiling technologies and cross-platform comparison. Next-generation microRNA expression profiling technology. Methods and protocols *2012*, 19–31.
13. Motameny, S., Wolters, S., Nürnberg, P., and Schumacher, B. (2010). Next generation sequencing of miRNAs-strategies, resources and methods. Genes *1*, 70–84.
14. Kim, M.A., Jung, E.J., Lee, H.S., Lee, H.E., Jeon, Y.K., Yang, H.K., and Kim, W.H. (2007). Evaluation of HER-2 gene status in gastric carcinoma using immunohistochemistry, fluorescence in situ hybridization, and real-time quantitative polymerase chain reaction. Hum. Pathol. *38*, 1386–1393.
15. Pall, G.S., and Hamilton, A.J. (2008). Improved northern blot method for enhanced detection of small RNA. Nat. Protoc. *3*, 1077–1084.
16. Wang, Y., Zhai, Y., Ding, Y., and Zou, Q. (2023). SBSM-Pro: support bio-sequence machine for proteins. Preprint at arXiv. https://doi.org/10.48550/arXiv:2308.10275.
17. Feng, H., Jin, D., Li, J., Li, Y., Zou, Q., and Liu, T. (2023). Matrix reconstruction with reliable neighbors for predicting potential MiRNA–disease associations. Briefings Bioinf. *24*, bbac571.
18. Wu, H., Liang, Q., Zhang, W., Zou, Q., El-Latif Hesham, A., and Liu, B. (2022). iLncDA-LTR: Identification of lncRNA-disease associations by learning to rank. Comput. Biol. Med. *146*, 105605.
19. Zhou, Z., Du, Z., Wei, J., Zhuo, L., Pan, S., Fu, X., and Lian, X. (2023). MHAM-NPI: Predicting ncRNA-protein interactions based on multi-head attention mechanism. Comput. Biol. Med. *163*, 107143.
20. Liu, W., Lin, H., Huang, L., Peng, L., Tang, T., Zhao, Q., and Yang, L. (2022). Identification of miRNA-disease associations via deep forest ensemble learning based on autoencoder. Briefings Bioinf. *23*, bbac104.
21. Wei, J., Zhuo, L., Zhou, Z., Lian, X., Fu, X., and Yao, X. (2023). GCFMCL: predicting miRNA-drug sensitivity using graph collaborative filtering and multi-view contrastive learning. Briefings Bioinf. *24*, bbad247.
22. Wang, C.C., Chen, X., Qu, J., Sun, Y.Z., and Li, J.Q. (2019). RFSMMA: a new computational model to identify and prioritize potential small molecule-miRNA associations. J. Chem. Inf. Model. *59*, 1668–1679.
23. Luo, J., Shen, C., Lai, Z., Cai, J., and Ding, P. (2021). Incorporating clinical, chemical and biological information for predicting small molecule-microRNA associations based on non-negative matrix factorization. IEEE ACM T. Comput. Bi. *18*, 2535–2545.
24. Ni, J., Cheng, X., Ni, T., and Liang, J. (2022). Identifying SM-miRNA associations based on layer attention graph convolutions network and matrix decomposition. Front. Mol. Biosci. *9*, 1009099.

25. Peng, L., Tu, Y., Huang, L., Li, Y., Fu, X., and Chen, X. (2022). DAESTB: inferring associations of small molecule-miRNA via a scalable tree boosting model based on deep autoencoder. Briefings Bioinf. *23*, bbac478.

26. Wang, S.H., Wang, C.C., Huang, L., Miao, L.Y., and Chen, X. (2022). Dual-network collaborative matrix factorization for predicting small molecule-miRNA associations. Briefings Bioinf. *23*, bbab500.

27. Wang, S., Liu, T., Ren, C., Wu, W., Zhao, Z., Pang, S., and Zhang, Y. (2023). Predicting potential small molecule-miRNA associations utilizing truncated schatten p-norm. Briefings Bioinf. *24*, bbad234.

28. Wang, C.C., Zhu, C.C., and Chen, X. (2022). Ensemble of kernel ridge regression-based small molecule-miRNA association prediction in human disease. Briefings Bioinf. *23*, bbab431.

29. Guan, N.N., Sun, Y.Z., Ming, Z., Li, J.Q., and Chen, X. (2018). Prediction of potential small molecule-associated micrornas using graphlet interaction. Front. Pharmacol. *9*, 1152.

30. Chen, X., Zhou, C., Wang, C.C., and Zhao, Y. (2021). Predicting potential small molecule-miRNA associations based on bounded nuclear norm regularization. Briefings Bioinf. *22*, bbab328.

31. Hamilton, W., Ying, Z., and Leskovec, J. (2017). Inductive representation learning on large graphs. Advances in Neural Information Processing Systems *30*.

32. Kipf, T.N., and Welling, M. (2017). Semi-supervised classification with graph convolutional networks. In International Conference on Learning Representations.

33. Xu, K., Hu, W., Leskovec, J., and Jegelka, S. (2018). How powerful are graph neural networks? In International Conference on Learning Representations.

34. Kurt, C.C., Bagca, B.G., Gunel, N.S., Birden, N., Shademan, B., Sogutlu, F., Ozates, N.P., Avci, C.B., Ozates, N.P., and Avci, C.B. (2021). Effect of valproic acid on miRNAs affecting histone deacetylase in a model of anaplastic thyroid cancer. Mol. Biol. Rep. *48*, 6085–6091.

35. Olde Loohuis, N.F.M., Kole, K., Glennon, J.C., Karel, P., Van der Borg, G., Van Gemert, Y., Van den Bosch, D., Meinhardt, J., Kos, A., Shahabipour, F., et al. (2015). Elevated microRNA-181c and microRNA-30d levels in the enlarged amygdala of the valproic acid rat model of autism. Neurobiol. Dis. *80*, 42–53.

36. Fu, M., Zhu, Y., Zhang, J., Wu, W., Sun, Y., Zhang, X., Tao, J., and Li, Z. (2021). MicroRNA-221-3p suppresses the microglia activation and seizures by inhibiting of HIF-1α in valproic acid-resistant epilepsy. Front. Pharmacol. *12*, 714556.

37. Luoni, A., and Riva, M.A. (2016). MicroRNAs and psychiatric disorders: From aetiology to treatment. Pharmacol. Ther. *167*, 13–27.

38. Jia, Z., Wang, K., Zhang, Y., Duan, Y., Xiao, K., Liu, S., and Ding, X. (2021). Icariin ameliorates diabetic renal tubulointerstitial fibrosis by restoring autophagy via regulation of the miR-192-5p/GLP-1R pathway. Front. Pharmacol. *12*, 720387.

39. Calura, E., Fruscio, R., Paracchini, L., Bignotti, E., Ravaggi, A., Martini, P., Sales, G., Beltrame, L., Clivio, L., Ceppi, L., et al. (2013). MiRNA landscape in stage I epithelial ovarian cancer defines the histotype specificities. Clin. Cancer Res. *19*, 4114–4123.

40. Li, J., Wu, R., Sun, W., Chen, L., Tian, S., Zhu, L., Meng, C., Zheng, Z., and Wang, W. (2023). What's Behind the Mask: Understanding Masked Graph Modeling for Graph Autoencoders. In Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, pp. 1268–1279.