Briefings in Bioinformatics, 2022, **23(6)**, 1–14

https://doi.org/10.1093/bib/bbac478

Advance access publication date 15 November 2022

Problem Solving Protocol

DAESTB: inferring associations of small molecule–miRNA via a scalable tree boosting model based on deep autoencoder

Li Peng, Yuan Tu, Li Huang, Yang Li, Xiangzheng Fu and Xiang Chen

Corresponding author. Li Peng, College of Computer Science and Engineering, Hunan University of Science and Technology, Xiangtan, 411201, Hunan, China. Email: plpeng@hnu.edu.cn

Abstract

MicroRNAs (miRNAs) are closely related to a variety of human diseases, not only regulating gene expression, but also having an important role in human life activities and being viable targets of small molecule drugs for disease treatment. Current computational techniques to predict the potential associations between small molecule and miRNA are not that accurate. Here, we proposed a new computational method based on a deep autoencoder and a scalable tree boosting model (DAESTB), to predict associations between small molecule and miRNA. First, we constructed a high-dimensional feature matrix by integrating small molecule–small molecule similarity, miRNA–miRNA similarity and known small molecule–miRNA associations. Second, we reduced feature dimensionality on the integrated matrix using a deep autoencoder to obtain the potential feature representation of each small molecule–miRNA pair. Finally, a scalable tree boosting model is used to predict small molecule and miRNA potential associations. The experiments on two datasets demonstrated the superiority of DAESTB over various state-of-the-art methods. DAESTB achieved the best AUC value. Furthermore, in three case studies, a large number of predicted associations by DAESTB are confirmed with the public accessed literature. We envision that DAESTB could serve as a useful biological model for predicting potential small molecule–miRNA associations.

Keywords: microRNA, small molecule, association prediction, deep autoencoder, scalable tree boosting

Introduction

With the development of sequencing technology, a large number of medical data has been accumulated in the biomedical field, which has provided researchers with more facilities to utilize the data to study the relationship between diseases and drugs. The presence of microRNA (miRNA), a short noncoding RNA, has been extensively studied in the human genome. More and more miRNAs have been discovered by researchers. Through intensive research, miRNAs have been found to play an important role in a variety of human life activities, influencing the proliferation, differentiation, aging and death of cells in humans and even other species [1]. Moreover, abnormal miRNA expression is closely associated with human diseases, which has led to the expectation that miRNAs can be used as clinical and prognostic biomarkers [2, 3]. An increasing number of researchers have worked on miRNA-targeting drugs, eventually discovering and demonstrating the potential of small molecule drugs to target miRNAs [4]. However, the use of traditional biological experiments to identify small molecule drug-associated miRNAs still suffers from the problem of blindness and requires significant experimental time and cost. Therefore, there is an urgent requirement to develop computational models to predict the associations between small molecule and miRNA. The computational models help the researchers to greatly improve experimental efficiency by allowing specific experimental validation of those associations that are most likely.

In recent decades, researchers have made great effort to exploit computational methods to solve different biological problems, such as the prediction of miRNA–disease associations [5–8], circular RNA–disease associations [9], gene–drug interactions [10] and so on. Chen et al. [11] have also developed a database for clinically or experimentally supported noncoding RNAs and drug targets associations. To a certain extent, these studies have contributed to the development of small molecule and miRNA associations prediction methods. In recent years, more and more researchers have proposed different computational models for predicting potential small molecule–miRNA associations [12–14]. Lv et al. [15] constructed a complete network by combining small molecule similarity network, miRNA similarity network and known small molecule-miRNA associations network. They calculated the similarity of small molecules and miRNAs separately by a weighted combination strategy, and then used the RWR (Random Walk With Restart) algorithm to predict the potential associations between small molecule and miRNA. SMiR-NBI [16] proposed a heterogeneous network based on drugs, miRNAs and genes using the NBI algorithm to predict miRNA responses to anti-cancer drugs. Later, TLHNSMMA [17] proposed a small molecule–miRNA associations

Li Peng is an associate professor at Hunan University of Science and Technology and Hunan Key Laboratory for Service computing and Novel Software. Her research interests are focused on machine learning, deep learning and bioinformatics.

Yuan Tu is a postgraduate student at Hunan University of Science and Technology. Her research interests include bioinformatics and machine learning.

Li Huang is a PhD student at Tsinghua University. His research interests include bioinformatics, complex networks and machine learning.

Yang Li is a PhD student at Xiangtan University. Her research interests include data mining and bioinformatics.

Xiangzheng Fu is a postdoctoral scholar at Hunan University. His research interest is classification of proteins in bioinformatics.

Xiang Chen is an assistant professor at Hunan University of Science and Technology. His research is in the areas of complex networks and bioinformatics.

Received: July 16, 2022. **Revised:** September 28, 2022. **Accepted:** October 8, 2022

prediction computational method based on a three-layer heterogeneous network. The heterogeneous network of this method contained not only small molecule similarity, miRNA similarity and known small molecule–miRNA associations, but also incorporated disease similarity, and known miRNA–disease associations, thus constructing a three-layer heterogeneous network. Information was then propagated through the heterogeneous network by constructing an iterative update algorithm to identify potential small molecule–miRNA associations. GISMMA [18] proposed Graphlet interaction-based predictive inference for small molecule–miRNA associations. This model used Graphlet interactions consisting of 28 isoforms to describe the relationship between two small molecules or between two miRNAs. The association scores between small molecule and miRNA were calculated by counting the number of small molecule–miRNA interactions in the small molecule similarity network and miRNA similarity network. HSSMMA [19] calculated the correlation between small molecules and miRNAs by a metric based on the search path of node-type sequences connecting two objects, and confirmed the association experimentally. RFSMMA [20] used a filter-based approach to extract reliable features of small molecules and miRNAs from integrated small molecule similarity and miRNA similarity, respectively. Machine learning techniques of random forest were then used to infer small molecule–miRNA associations. SNMFSMMA [21] proposed a new small molecule–miRNA associations prediction model, symmetric non-negative matrix decomposition, to predict the potential association of SM–miRNA pairs. BNNRSMMA [22] first defined a novel matrix to represent small molecule–miRNA heterogeneous networks using miRNA–miRNA similarity, small molecule–small molecule similarity and known small molecule–miRNA associations. This matrix was then completed by minimizing its nuclear parametric number, and alternating directional multiplication was used to minimize the nuclear parametric number and obtain the prediction scores. Among them, they introduced a regularization term to tolerate the noise in the integrated similarity. Wang et al. [23] proposed a nuclear ridge regression-based approach to predict the small molecule–miRNA associations. Feature subsets of small molecules and miRNAs were first constructed separately, and homogeneous base learners were trained according to the different feature subsets, and finally the average score obtained by these base learners was used as the SM–miRNA associations score. Wang et al. [24] proposed a new dual network collaborative matrix decomposition (DCMF) method to predict potential SM–miRNA associations. They firstly preprocessed the missing values of the SM–miRNA associations matrix using the WKNKN method, and then constructed a matrix decomposition model of the dual network to obtain the feature matrices containing the potential features of small molecules and miRNAs, respectively. Finally, the predicted SM–miRNA associations score matrix was gained by calculating the inner product of the two feature matrices. Most of the above methods are based on machine learning and matrix decomposition approaches. These models have all achieved encouraging results and have played an important role in the development of computational methods for small molecule–miRNA associations recognition. However, they have certain problems or limitations: the experimentally validated small molecule–miRNA associations are very limited, and there are many negative associations. Performed on such noisy and sparse small molecule–miRNA associations networks, the predictors tend to detect many false negative associations. This also motivated us to develop a new deep learning model to further reveal potential associations between small molecule–miRNA.

In this work, we proposed a scalable tree boosting (STB) model based on a deep autoencoder, named deep autoencoder and a scalable tree boosting model (DAESTB), to predict small molecule–miRNA associations. First, we constructed a high-dimensional feature matrix of small molecule–miRNA combined association pairs using small molecule similarity, miRNA similarity and known small molecule–miRNA associations to maximize the use of information from miRNAs and small molecules. Then, we utilized a three-layer deep autoencoder to perform dimensionality reduction on the constructed high-dimensional matrix to extract potential feature representation. This not only improved the learning efficiency of the model, but also eliminated noise. Finally, we used an STB method to predict potential associations between small molecule–miRNA. In the experiments, we performed 5-fold and 10-fold CV on the dataset 1 and the dataset 2, respectively, and compared them with other classifiers. The DAESTB model was shown to be more accurate than other classifiers in predicting the associations. In addition, we also compared them with other several methods, respectively, and the AUC accuracy of the DAESTB method was relatively higher. Finally, case studies were also deployed to demonstrate the ability of DAESTB in identifying candidate miRNAs which are potentially relevant to small molecules. The DAESTB calculation method has the following advantages: (i) the extraction of potential features of small molecule–miRNA association pairs by deep autoencoder can remove the noise in the data, thus effectively improving the prediction efficiency; (ii) STB is a scalable tree boosting method, and the parameters used in this paper are with all default values, so there is no need to adjust the parameters for the model to achieve good results. (iii) Comparing with relevant models and other machine learning classifiers, DAESTB achieves the best results in terms of area under receiver operating characteristic (ROC) curves, which indicates that the predictive performance of the DAESTB method is more accurate.

Materials

Known small molecule–miRNA associations

We obtained known small molecule–miRNA associations from SM2miR (version 1) [25], in which the total number of known associations was 664. A total number of 831 small molecules were then extracted and integrated from SM2miR, DrugBank [26] and PubChem [27], and 541 miRNAs were collected from SM2miR, HMDD [28], miR2Disease [29] and PhenomiR [30]. On the dataset 1, there are 831 small molecules, 541 miRNAs and 664 known small molecule–miRNA associations. Among them, only 39 small molecules and 286 miRNAs are associated, while added 792 completely new small molecules and 255 new miRNAs without any known association. Then, we removed the small molecules and miRNAs that had no association on the dataset 1, while retaining the 39 small molecules, 286 miRNAs and 664 known small molecule–miRNA associations that formed the dataset 2. We constructed an adjacency matrix $A \in \mathbb{R}^{nm \times ns}$ to represent the associations between small molecule and miRNA, where the i-th row indicates that the miRNA is $m(i)$ and the j-th column indicates the small molecule $s(j)$. If there is an association between $m(i)$ and $s(j)$, $A(i,j)$ is 1; otherwise, $A(i,j)$ is 0. The association matrix A is defined as

$$A(i,j) = \begin{cases} 1, & \text{if } m(i) \text{ and } s(j) \text{ has an association} \\ 0, & \text{otherwise} \end{cases}, \quad (1)$$

where $i \in \{1, 2, \ldots, nm\}$ represents miRNA $m(i)$, $j \in \{1, 2, 3, \ldots, ns\}$ represents small molecule $s(j)$, nm and ns represent the number of miRNAs and small molecules, respectively. In addition, the distribution map of miRNAs corresponding to small molecules are shown in the supplementary material (Figures S1–S2).

Integrated small molecule–small molecule similarity

The integrated small molecule similarity was derived by integrating four similarity calculation methods [15], which include the phenotypic similarity between diseases associated with small molecules [31], functional identity based on small molecule target gene set [32], similarity based on small molecule chemical structure [33] and similarity based on small molecule side effects [32].

Among them, disease-based phenotypic similarity was calculated using Jaccard score, which indicated that the more the diseases shared by small molecules, the greater the similarity between small molecules; functional congruence based on target gene set was calculated by extracting target genes of small molecules from DrugBank [34] and TTD [35] databases and based on GSFS [32] score, and if the target genes of two small molecules genes had functional identity, then they were more similar; chemical structure-based similarity was calculated by SIMCOMP [33], a graph-based approach that searched for the largest common subgraph isomorphism by finding the largest cluster in the association graph, thus responding to the global score of similarity; side-effect similarity based on small molecules was calculated using a Jaccard score in which the top 13 most relevant side-effects for each small molecule were extracted from the SIDER [36], and the more side-effects the two small molecules contained, the more similar they were. The final similarity was calculated with the following formula:

$$SMS = \frac{\alpha_1 S_{sm}^d + \alpha_2 S_{sm}^f + \alpha_3 S_{sm}^c + \alpha_4 S_{sm}^s}{\alpha_1 + \alpha_2 + \alpha_3 + \alpha_4}, \tag{2}$$

where S_{sm}^d, S_{sm}^f, S_{sm}^c and S_{sm}^s represent four types of similarity, i.e. similarity based on denotational phenotype, similarity based on functional consistency, similarity based on chemical structure and similarity based on side effects; $\alpha_i (i \in \{1, 2, 3, 4\})$ is the weight parameter to measure the contributions of each similarity, and they are all set to 1 in order to reduce the bias of each similarity metric.

Integrated miRNA–miRNA similarity

We obtained miRNA similarity by integrating the identification of phenotypic similarity between miRNA-related diseases [31] and functional identity based on the set of target genes [32].

Similarly, disease-related miRNAs were extracted from three databases, HMDD [36], miR2Disease [25] and PhenomiR [30], and the Jaccard score was used to define the similarity between miRNAs. If two miRNAs are associated with more diseases in common, the more similar they were. And the target gene-based functional identity was calculated as the similarity between two miRNAs by calculating the functional identity between any two miRNA target gene sets, which was calculated by GCFS [32] score; the higher the GCFS score between target gene sets, the stronger the similarity between miRNAs. Eventually, we obtained the integrated miRNA similarity as the following formula:

$$MMS = \frac{\beta_1 S_{mm}^d + \beta_2 S_{mm}^f}{\beta_1 + \beta_2}, \tag{3}$$

where S_{mm}^d denotes the phenotype-based indicated similarity and S_{mm}^f denotes the functional consistency-based similarity; similarly, $\beta_j (j \in \{1, 2\})$ is the weight of each similarity measure, and they are set to 1 in order to reduce the bias of each similarity metric.

Methods

The proposed DAESTB model includes the following three steps, which is illustrated in Figure 1. In the first step, we preprocessed the data of three matrices, namely, small molecule–small molecule similarity, miRNA–miRNA similarity and known small molecule–miRNA associations, and constructed them into a high-dimensional small molecule–miRNA association pairs feature matrix F_{MS}. In detail, small molecule–miRNA association pairs consist of all small molecule and miRNA paired with each other. In the second step, the constructed high-dimensional feature matrix F_{MS} was processed using the deep autoencoder by reducing feature dimensionality, thus obtaining the corresponding low 128-dimensional feature representation matrix F'_{MS}. In the third step, the generated low-dimensional features were used to train an STB model to predict the potential associations between small molecule and miRNA.

Construction of feature representation matrix

In this section, we mainly constructed a new high-dimensional feature matrix of small molecule–miRNA association pairs by integrating three different dimensional matrices, i.e. small molecule–small molecule similarity matrix SMS, miRNA–miRNA similarity matrix MMS and known small molecules and miRNA associations matrix Y. The high-dimensional matrix had $(ns \times nm)$ rows and $(ns + nm)$ columns; each row represented a small molecule–miRNA association pair, and each column represented the feature presentation in the small molecule–miRNA association pair. Considering from the perspective of small molecules, we could find that in each small molecule–miRNA association pair, there existed other small molecules similar to the corresponding small molecules in the association pair, and they could form an association pairs feature matrix $F_{SM} \in R^{(ns \times nm) * ns}$ associated with small molecules.

Similarly, considering from the perspective of miRNAs, for each small molecule–miRNA association pair, there existed other miRNAs similar to the corresponding miRNA in the association pair, which could form an miRNA-related association pairs feature matrix $F_{miRNA} \in R^{(ns \times nm) * nm}$.

We connected these two association pairs feature matrices according to the corresponding association pair in each row, and finally we could get a more comprehensive small molecule–miRNA association pairs feature matrix $F_{MS} \in R^{(ns \times nm) * (ns + nm)}$, as shown in Figure 1. The number of small molecule is ns, the number of miRNA is nm and a total of $(ns \times nm)$ pairs existed for small molecule–miRNA combinations, regardless of whether small molecule was associated with miRNA or not.

Using deep autoencoder to extract low-dimensional feature

In the previous section, we obtained a high-dimensional feature representation of small molecule–miRNA association pairs, $(nm+ns)$-dimensional. High-dimensional features are particularly time-consuming in computation and also affect the accuracy of prediction. Therefore, we used a deep autoencoder with three hidden layers to reduce the feature dimension of the constructed

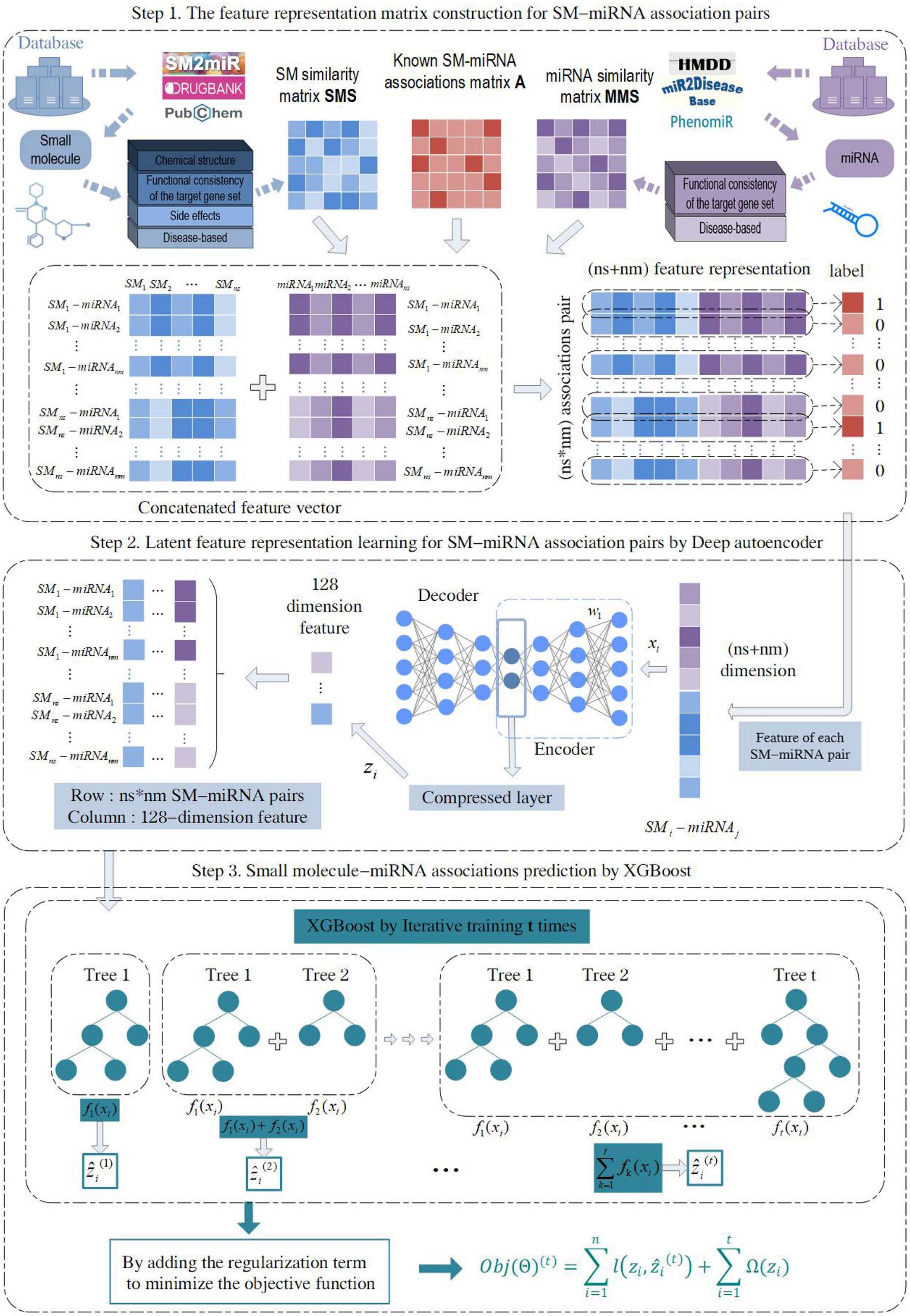
Step 1. The feature representation matrix construction for SM–miRNA association pairs
Database
SM2miR
DRUGBANK
PubChem
Small molecule
Chemical structure
Functional consistency of the target gene set
Side effects
Disease-based
SM similarity matrix SMS
Known SM-miRNA associations matrix A
miRNA similarity matrix MMS
HMDD
miR2Disease Base
PhenomiR
Database
miRNA
Functional consistency of the target gene set
Disease-based
SM_1 SM_2 ··· SM_{ns}
$miRNA_1$ $miRNA_2$ ··· $miRNA_{nz}$
$SM_1-miRNA_1$
$SM_1-miRNA_2$
$SM_1-miRNA_{nm}$
$SM_{ns}-miRNA_1$
$SM_{ns}-miRNA_2$
$SM_{ns}-miRNA_{nm}$
Concatenated feature vector
(ns+nm) feature representation
label
(ns*nm) associations pair
1
0
0
0
1
0
Step 2. Latent feature representation learning for SM–miRNA association pairs by Deep autoencoder
$SM_1-miRNA_1$
$SM_1-miRNA_2$
$SM_1-miRNA_{nm}$
$SM_{ns}-miRNA_1$
$SM_{ns}-miRNA_2$
$SM_{ns}-miRNA_{nm}$
Row : ns*nm SM–miRNA pairs
Column : 128–dimension feature
128 dimension feature
z_i
Decoder
w_1
x_i
Encoder
Compressed layer
(ns+nm) dimension
Feature of each SM–miRNA pair
$SM_i-miRNA_j$
Step 3. Small molecule–miRNA associations prediction by XGBoost
XGBoost by Iterative training t times
Tree 1
Tree 1
Tree 2
Tree 1
Tree 2
Tree t
$f_1(x_i)$
$\hat{z}_i^{(1)}$
$f_1(x_i)$
$f_2(x_i)$
$f_1(x_i)+f_2(x_i)$
$\hat{z}_i^{(2)}$
$f_1(x_i)$
$f_2(x_i)$
···
$f_t(x_i)$
$\sum_{k=1}^{t} f_k(x_i)$
$\hat{z}_i^{(t)}$
By adding the regularization term to minimize the objective function
$Obj(\Theta)^{(t)} = \sum_{i=1}^{n} l(z_i, \hat{z}_i^{(t)}) + \sum_{i=1}^{t} \Omega(z_i)$

Figure 1. The flowchart of DAESTB. Step 1, a high-dimensional feature matrix of small molecule–miRNA association pairs is constructed; Step 2, feature dimension of small molecule–miRNA association pairs is educed by deep autoencoder to obtain potential features of association pairs; Step 3, XGBoost is used to predict potential associations between small molecule and miRNA.

high-dimensional small molecule–miRNA association pairs feature matrix and extract the most essential features of each association pair.

Specifically, the deep autoencoder was trained to extract a low-dimensional representation of features from small molecule–miRNA association pairs. The autoencoder learned the function as shown in equation (4) by

$$\hat{X} = \varphi_{w_1,b_1}(X) \approx X, \tag{4}$$

where the input vector is X, w_1 denotes the weights and b_1 denotes the bias. The model training process in the deep autoencoder is divided into encoding and decoding. During encoding, the autoencoder used the mapping function $\psi(\cdot)$ to compress the input X (high-dimensional feature representation of small molecule–miRNA association pairs) into a low-dimensional representation vector according to equation (5):

$$h = g_{\theta_1}(X) = \psi\left(w_1 X + b_1\right), \tag{5}$$

where

$$\psi(x) = \frac{1}{1 + \exp(-x)} \tag{6}$$

and $\psi(\cdot)$ denotes the encoding function and $h = g_{\theta_1}(X)$ is the low-dimensional representation obtained by encoding. In the decoding process, the obtained low-dimensional representation h was decoded into features $\hat{X}$ which were similar to the input feature representation:

$$\hat{X} = g_{\theta_2}(h) = \phi\left(w_2 h + b_2\right), \tag{7}$$

where w_2 denotes the weight and b_2 denotes the bias; $\phi(\cdot)$ is the decoding function, and $\hat{X}$ is the recovered feature representation. During the iterative process, the reconstruction error loss function was defined as

$$\mathcal{L}(X, \hat{X}) = \sum_{k=1}^{K} \|\hat{X} - X\|^2, \tag{8}$$

where $k \in K$ denotes the number of training samples. According to the literature [37], we set the compressed feature dimension of the autoencoder to 128. During the training process, we used mean square error (MSE) as the loss function and Adam as the optimization algorithm to reduce the loss. For the dataset 1, the high-dimensional feature matrix of the combined small molecule and miRNA association pairs is (831+541) dimensions, i.e. 1372 dimensions, while for the dataset 2, the high-dimensional feature matrix is (39+286) dimensions, i.e. 325 dimensions. In order to achieve the effect of dimensionality reduction, the number of neurons in the hidden layer must be set smaller than the number of neurons in the input layer, while the number of neurons in the output layer needs to be set equal to the number of neurons in the input layer. Here, we chose 512, 256 and 128, respectively, for the number of neurons in these three layers on the dataset 1, thus achieving the reduction of the high dimensionality to 128 dimensions, while on the dataset 2, the number of neurons in these three layers was chosen to be 250, 200 and 128, respectively. Ultimately, we obtained the low-dimensional feature matrix $F_{MS}{}'$ of small molecule–miRNA association pairs.

Predicting potential associations by STB

We used an STB [38] method (XGBoost) to predict potential associations of small molecule and miRNA. We called the STB method simply STB. STB is similar to Boosting, which iterates through a gradient of multiple weak classifiers and then obtains accurate classification. However, the weak classifiers are not independent of each other, and the next classifier is obtained by optimizing the results of the previous classifier. STB is a new tree-learning algorithm for processing sparse data, which has been well used in miRNA–disease associations prediction [39]. In our study, there are only a few known associations between small molecule and miRNA, and the association matrix they construct is very sparse, which allows us to make better use of this method for predicting potential associations. The low-dimensional feature matrix $F_{MS}{}'$ is fed into the STB model of the obtained small molecule–miRNA association pairs and train multiple weak classifiers to obtain the best gradient regression tree, as shown in the following equation (9):

$$\begin{aligned}
\hat{z}_i^{(0)} &= 0 \\
\hat{z}_i^{(1)} &= f_1(x_i) = \hat{z}_i^{(0)} + f_1(x_i) \\
\hat{z}_i^{(2)} &= f_1(x_i) + f_2(x_i) = \hat{z}_i^{(1)} + f_2(x_i) \\
&\cdots \\
\hat{z}_i^{(t)} &= \sum_{k=1}^{t} f_k(x_i) = \hat{z}_i^{(t-1)} + f_t(x_i),
\end{aligned} \tag{9}$$

where $\hat{z}_i^{(j)}$ denotes the classification result of the jth classifier, $j \in (1, 2, 3, \ldots, t)$ and t is the number of weak classifiers in the model; $f_k \epsilon \mathcal{F}$ denotes the space of regression trees (i.e. CART). By minimizing the following objective function:

$$\begin{aligned}
\mathrm{Obj}(\Theta)^{(t)} &= \sum_{i=1}^{n} l\left(z_i, \hat{z}_i^{(t)}\right) + \sum_{i=1}^{t} \Omega(z_i) \\
&= \sum_{i=1}^{n} l\left(z_i, \hat{z}_i^{(t-1)} + f_t(x_i)\right) + \sum_{i=1}^{t} \Omega(z_i),
\end{aligned} \tag{10}$$

where $\mathcal{L}(\Theta) = \sum_{i=1}^{n} l\left(z_i, \hat{z}_i^{(t)}\right)$ is the loss function used to calculate the error of the training set, and $\Omega(\Theta) = \sum_{i=1}^{t} \Omega(z_i)$ is the regularization term used to penalize the complexity of the model to avoid overfitting. By minimizing the objective function through iterative training, new small molecule–miRNA features can be fitted effectively, and the potential associations between small molecule and miRNA can be predicted finally.

Results

Performance evaluation

In this work, 5-fold cross-validation (5-fold CV) and 10-fold cross-validation (10-fold CV) were utilized on the dataset 1 and the dataset 2, respectively, to evaluate the predictive performance of our method. All predicted small molecule–miRNA pairs were ranked according to the obtained scores. Based on the rankings, we used the receiver operating characteristic (ROC) curves to illustrate the performance of our model in the two cross-validations. Our AUC values were 0.9863 and 0.9856 for the dataset 1 by 5-fold and 10-fold CV, respectively, and 0.8653 and 0.8812 for the dataset 2, as shown in Figure 2. We further evaluated several commonly measured metrics, namely accuracy (ACC), precision

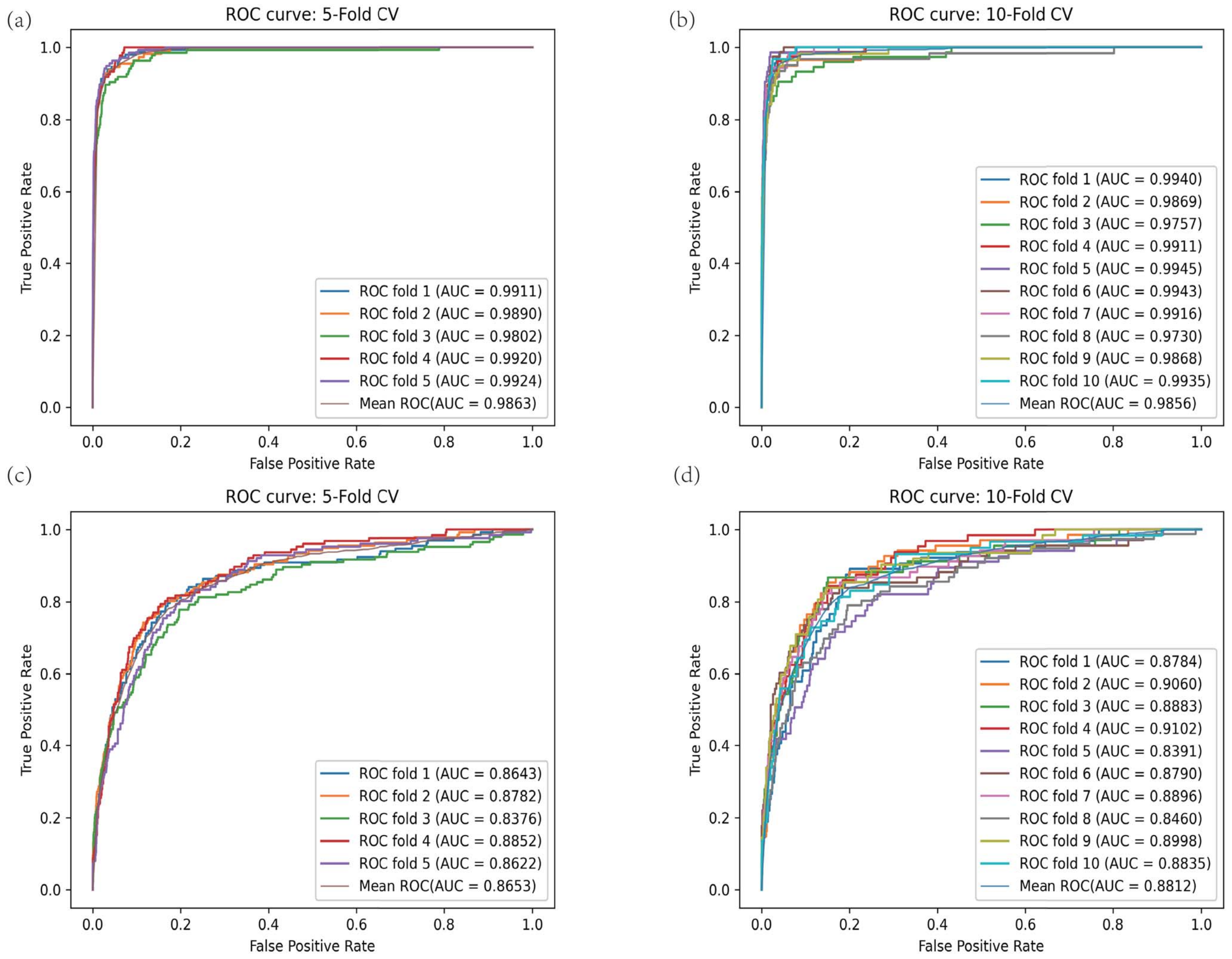

Figure 2. (A) DAESTB performs 5-fold CV on the dataset 1; **(B)** DAESTB performs 10-fold CV on the dataset 1; **(C)** DAESTB performs 5-fold CV on the dataset 2; **(D)** DAESTB performs 10-fold CV on the dataset 2.

(PRE), sensitivity, F1-score (F1), Mathews correlation coefficient (MCC) and the area under the precision recall curve (AUPRC), calculated by the following equations:

$$\mathrm{ACC} = \frac{TP + TN}{TP + TN + FP + FN} \tag{11}$$

$$\mathrm{PRE} = \frac{TP}{TP + FP} \tag{12}$$

$$\text{Recall } = \frac{TP}{TP + FN} \tag{13}$$

$$\mathrm{F1} = \frac{2 \times \text{ Precision } \times \text{ Recall}}{\text{Precision } + \text{ Recall}} \tag{14}$$

$$\text{Sensitivity } = \frac{TP}{TP + FN} \tag{15}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{16}$$

Comparison on different potential characteristic dimensions

In the DAESTB method, we used the deep autoencoder to perform dimensionality reduction on the constructed high-dimensional small molecule–miRNA combination association pairs, so as to obtain the potential features of the association pairs. To illustrate the dimensionality of the features after the deep autoencoder dimensionality reduction, we set the dimensions to 8, 16, 32, 64, 128 and 256 on the dataset 1, respectively, and compared them by 5-fold CV to find out the dimensionality that achieved the best results of the model. From Figure 3, it can be shown that the best performance during the model training is the best when the dimensionality reached 128. In detail, we used AUC as the most important evaluation metric. Therefore, we set the number of the potential feature dimension of the combined small molecule–miRNA association pairs as 128.

Parameter setting

Layers of deep autoencoder

A high-dimensional feature matrix of small molecule–miRNA combinatorial association pairs was trained by a three-layer deep

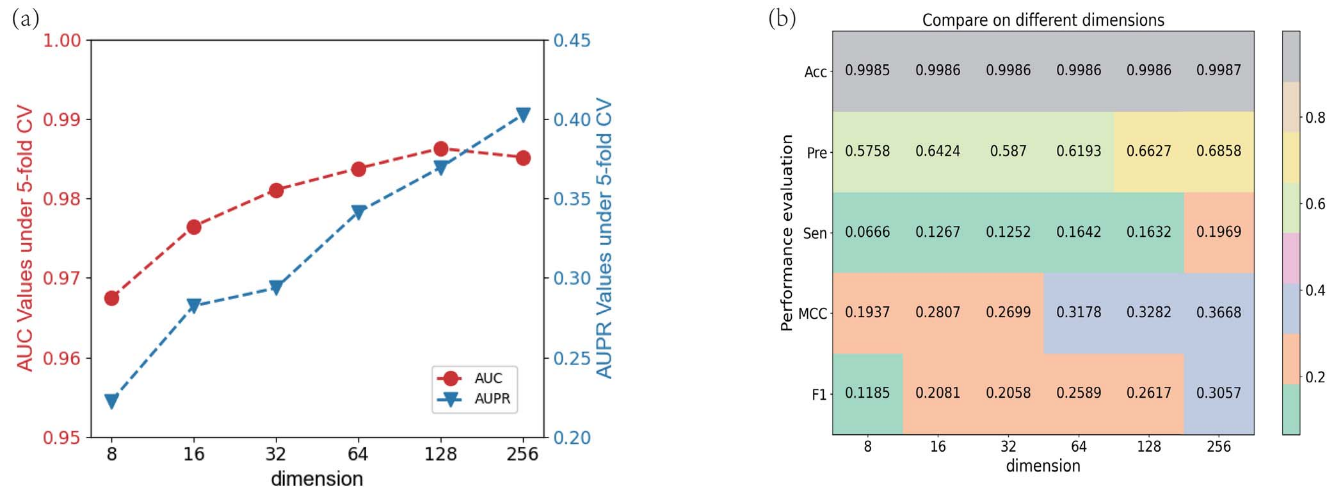

Figure 3. (A) AUC and AUPR values achieved by DAESTB, with varying the latent dimension of the last layer in the deep autoencoder model from 8 to 256; **(B)** DAESTB was compared with other dimensions on the dataset 1 by 5-fold CV for other evaluation performance metrics.

autoencoder, and eventually low-dimensional latent features are extracted. The number of layers l of the deep autoencoder was determined from $\{2, 3, 4\}$ on the dataset 1 by 5-fold CV. As can be seen in Figure 4(A), the value of 0.9863 for the AUC is the highest when $l = 3$.

Deep autoencoder loss in extracting low-dimension features

In this process, we presented the deep autoencoder training results in Figure 4(B) and 4(C). As the epoch number increases, the value of the loss function of DAESTB first decreased rapidly, then decreased gently, and eventually reached a steady state. Hence, we set epoch to 20 and batch size to 128 on the dataset 1 by 5-fold CV.

Adjustment of learning rates

In this section, we adjusted the parameters of DAESTB. In the XGBoost predictive method, the learning rate γ was determined from $\{0.1, 0.3, 0.5, 0.7\}$ and the rest of the parameters were defaulted on the dataset 1 by 5-fold CV. The purpose of adjusting the parameter γ is to make the prediction of XGBoost more accurate. The training results are shown in Figure 4(D). DAESTB model has the highest prediction accuracy when γ is 0.3, and its corresponding AUC value is 0.9863. Assuming that all parameters are set with default values, the AUC value is still 0.9863.

Comparison with other dimensionality reduction methods

To evaluate the performance of DAESTB, we compared it with six other dimensionality reduction methods, including Principal Component Analysis (PCA) [40], t-Distributed Stochastic Neighbor Embedding (TSNE) [41], Uniform Manifold Approximation and Projection (UMAP) [42], Factor Analysis (FA) [43], Independent Component Correlation Algorithm (ICA) [44] and FastICA [45]. To be fair, all the dimensionality reduction methods were experimented on the dataset 1 using 5-fold CV, and the final AUC values of the models were calculated. After experimental validation, we can see from Figure 4(E) that the AUC values for DAESTB, PCA, TSNE, UMAP, FA, FastICA and ICA on the dataset 1 are 0.9863, 0.9855, 0.9584, 0.9825, 0.9847, 0.9837 and 0.9854, respectively. The results show that DAESTB has the highest AUC value, which indicates that the deep autoencoder in DAESTB is better than other dimensionality reduction methods.

Comparison with other classifiers

To further evaluate the performance of DAESTB, we compared it with seven other classifiers, including logistic regression (LR) [46], Naive Bayes (Bayes) [47], decision tree (DT) [48], AdaBoost [49], random forest (RF) [50], gradient boosting (GB) [51], support vector machine (SVM) [52], LightGBM (LGBM) [53] and CatBoost (CB) [54]. To be considered fair, the same data were used for all classifiers and the final ROC curves were drawn using 5-fold and 10-fold CV for these eight classifiers on the dataset 1 and the dataset 2, respectively. In addition, other performance evaluation metrics for these eight classifiers are shown in the supplementary material (Tables S1–S2).

After experimental validation, we can see from Figure 4(F) and 4(G) that the AUC value of the DAESTB method is the highest on the dataset 1 and the dataset 2, which also indicates that the DAESTB classifier has better performance than other classifiers. The reasons are assumed as follows: (1) Although logistic regression is faster to train, it cannot solve the nonlinearity problem and the accuracy is not very high. (2) In theory, the Naive Bayes model has less error compared with other classifiers, but this is based on an assumption that attributes of the Naive Bayes model are independent of each other. However, in our data, the attributes are not independent of each other, so the classification is not as effective. (3) The common decision tree approach tends to produce an overly complex model, which can have a poor generalization performance to the data and can lead to overfitting. Some features are difficult to be learnt by the decision tree, thus causing bias. (4) Random forest is more resistant to overfitting, but when there is noise in the sample data, it will affect the prediction results. (5) SVMs consume space mainly by storing training samples and kernel matrices. Since SVMs solve support vectors with the help of quadratic programming, the training time is relatively long. In addition, its performance depends mainly on the selection of the kernel function. (6) AdaBoost, Gradient Boosting and STB all belong to special integrated learning methods. In AdaBoost model training, the execution depends on the selection of weak classifiers and is sample sensitive. In Gradient Boosting training, there is a serial relationship between the base learners, making it difficult to train data in parallel. While traditional Gradient Boosting uses CART trees as base learners, STB further supports linear classifiers with a regular term added to the cost function for controlling the complexity of the model. (7) LightGBM, simply

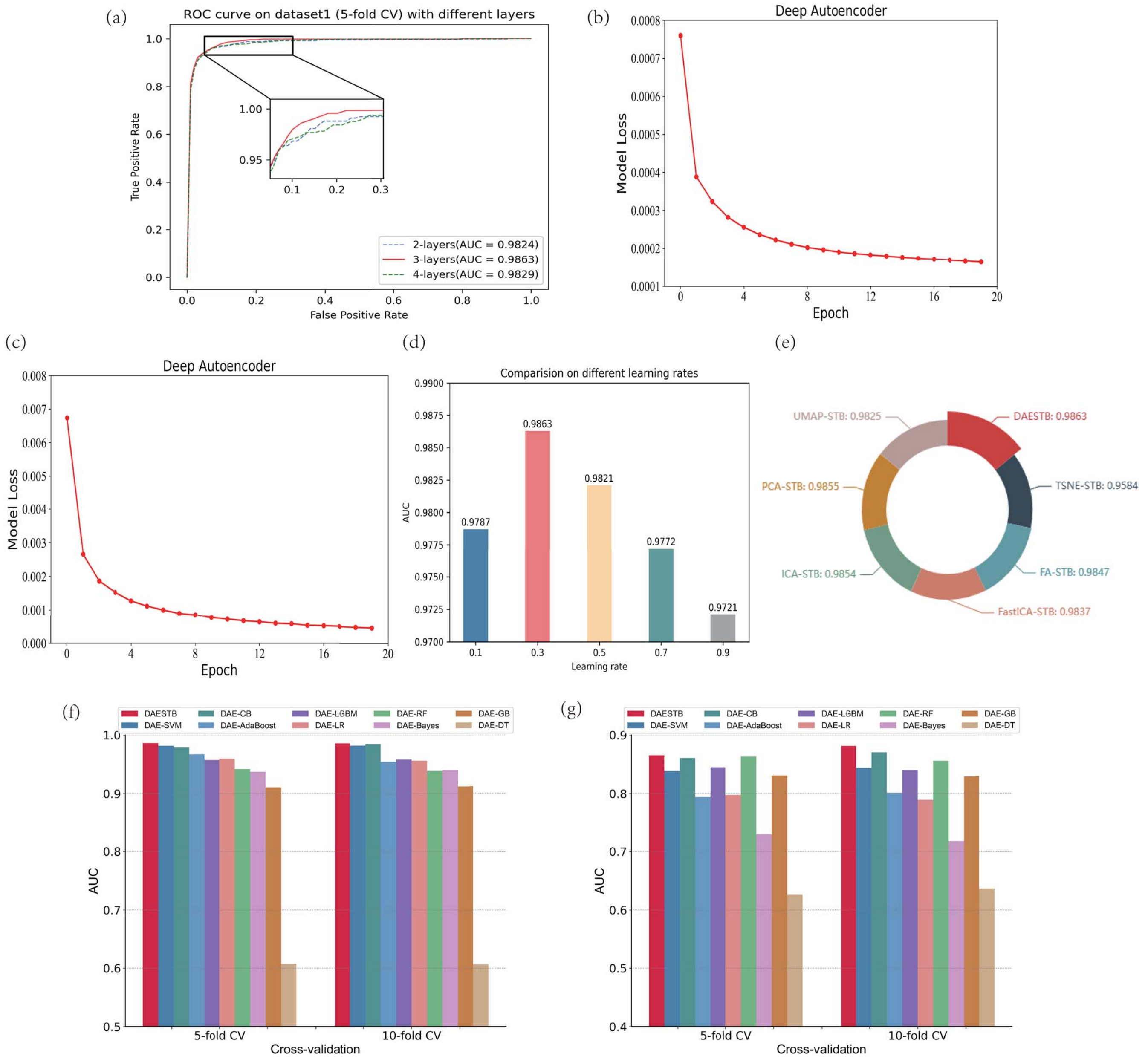

Figure 4. (A) Comparison of the number of layers in deep autoencoder; **(B)** Loss curve of DAESTB on the dataset 1 when extracting potential features using deep autoencoder; **(C)** Loss curve of DAESTB on the dataset 2 when extracting potential features using deep autoencoder; **(D)** The performance comparison of DAESTB with different learning rate in terms of AUC. **(E)** Comparison of DAESTB with six other methods of dimensionality reduction; **(F)** The performance comparison of DAESTB with other classifiers in terms of AUC, with 5-fold CV and 10-fold CV experiments on the dataset 1. DAESTB achieved the highest AUC value; **(G)** The performance comparison of DAESTB with other classifiers in terms of AUC, with 5-fold CV and 10-fold CV experiments on the dataset 2.

LGBM, is a bias-based algorithm which is more sensitive to noise. After adjusting the parameters, the learning rate was set to 0.01 and number of weak classifiers was set to 100, which resulted in the best AUC value. (8) CatBoost, shortly referred to as CB, requires a lot of memory and time for the processing of category-based features, and the setting of different random numbers has an impact on the model prediction results. After adjusting the parameters, we finally chose the best AUC value when the iterations were 1000 and the learning rate was 0.5. Finally, in this experiment, the parameters used by STB were default values, and there was no need to adjust the parameters for the model to achieve good results, which is one of the major advantages.

Comparison with other methods

To further evaluate the ability of the DAESTB model to predict potential small molecule–miRNA associations, we selected four other methods, namely, EKRRSMMA [23], GISMMA [18], BNNRSMMA [22], RWR [15], and compared them with our method by 5-fold CV on the dataset 1 and the dataset 2, respectively. For the four comparison methods, we used the same parameter settings as those set in the original paper, which are shown in the supplementary material (Table S3). Figure 5 demonstrates the experimental results of DAESTB with the other four competitive methods. On the dataset 1, the AUC values of DAESTB, EKRRSMMA, GISMMA, BNNRSMMA and RWR were

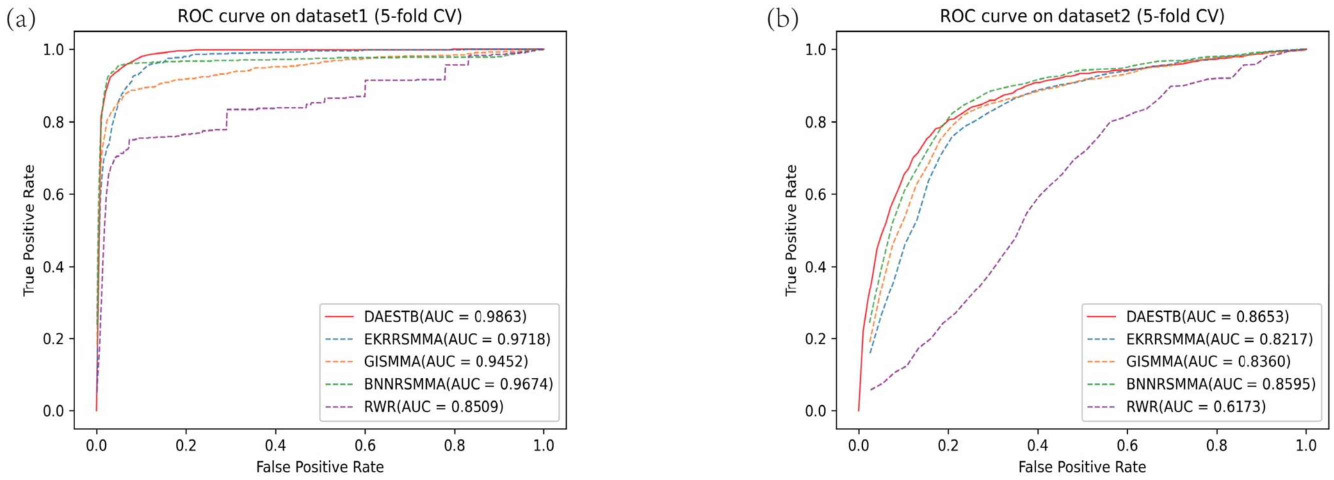

Figure 5. (A) The ROC curve of DAESTB on the dataset 1 compared with other methods; **(B)** The ROC curve of DAESTB on the dataset 2 compared with other methods.

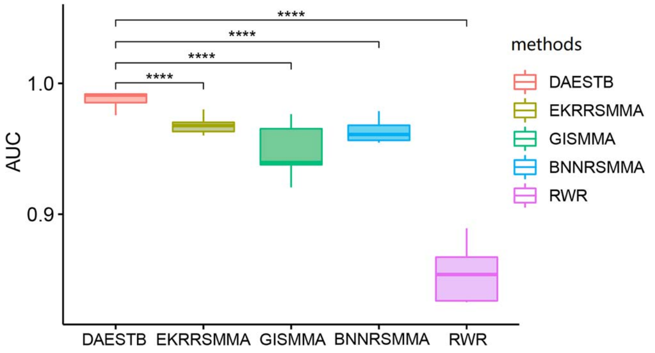

Figure 6. DAESTB significantly outperformed other methods in terms of AUC. (paired t-test, **** stands for *P*-value¡0.0001).

0.9863, 0.9718, 0.9452, 0.9674, and 0.8509, respectively, with the AUC of DAESTB being 0.0117 higher than that of the second best model (EKRRSMMA); on the dataset 2, the AUC values were 0.8653, 0.8217, 0.8360, 0.8595 and 0.6173, respectively, where the AUC of our method was 0.0042 higher than the second best model (BNNRSMMA). By comparing these two datasets, it is ultimately evident that DAESTB wins the best performance in predicting potential associations of small molecule–miRNA. In addition, we assessed the performance of the DAESTB method against the other four methods EKRRSMMA, GISMMA and BNNSMMA with 10 times 5-fold CV and 50 AUC values were collected for all methods. The AUC values of DAESTB were then statistically tested against those of EKRRSMMA, GISMMA and BNNSMMA using paired t-test, and the significant differences between DAESTB and other methods were marked. As shown in Figure 6, we can conclude that DAESTB does work more effectively than the other methods.

Case Studies

To evaluate the performance of DAESTB in predicting potential associations new small molecule and miRNA, we further implemented case studies for three common small molecules named 17beta-estradiol, 5-aza-2'-deoxycytidine, doxorubicin. It is worth noting that we predict unknown associations using all known small molecule–miRNA associations on the dataset 1. After employing the DAESTB model, we were able to obtain the score in the row corresponding to each small molecule in association matrix. Then, we ranked the association candidates based on the predicted association scores. Finally, we selected the top 50 corresponding rows of the studied small molecules and validated the associations between small molecule and miRNA through the experimental literature in the PubMed database.

The first small molecule is 17β-estradiol, also known as E2, which is a human estrogen with important effects in human reproduction [55]. After the DAESTB experiment, we obtained a ranking of all miRNAs associated with 17β-estradiol. 11 of the top 20 miRNAs were validated, while 24 of the top 50 miRNAs were validated (Table 1). For example, knockdown of XIST resulted in inhibition of cell proliferation and upregulation of miR-29c after treatment with a combination of 17β-estradiol and progesterone using a 3D sphere culture system [56]. Small RNA-Seq analysis of 17β-estradiol-treated muscles versus controls identified 36 differentially expressed miRNAs, with some important myogenic miRNAs, such as miR-145, being down-regulated [57]. In [58], to explore whether miR-155 is involved in E2 regulation of estrogen-responsive gene expression, miR-155 expression in human breast cancer cells was assessed by real-time PCR. Treatment of MCF-7 cells with E2 increased miR-155 expression promoted proliferation and reduced apoptosis. In assessing the expression levels of miR-34a and miR-15b in a transgenic model of HPV16K14E7 (K14E7) in mice after chronic estrogen (E2) treatment and their involvement in cervical cancer, although miR-34a expression was elevated by the HPVE7 oncogene, this expression was downregulated in the presence of both the E7 oncoprotein and chronic E2 in cervical cancer [59].

The second small molecule is 5-aza-2'-deoxycytidine (5-Aza-CdR), which is a nitrogenous nucleoside analog used in cell culture that re-expresses certain genes [60]. After the DAESTB experiment, we obtained all the rankings of miRNAs associated with 5-Aza-CdR. 11 of the top 20 miRNAs were validated, while 24 of the top 50 miRNAs were validated (Table 2). For example, human adult DPSCs were isolated from normal third molars using 5-Aza-2'-deoxycytidine, and miR-143 expression was significantly decreased during induced myogenic DPSCs [61]. Following 5-aza-2'-deoxycytidine treatment of gastric cancer cells, miR-10b methylation was significantly reduced and the expression of miR-10b and HOXD4 1 kb downstream of miR-10b was greatly restored [62]. 5-Aza-CdR induced cell death and increased miR-146a expression in LNCaP and PC3 cells, where miR-146a

Table 1. Based on the dataset 1, the top 50 miRNAs associated with the small molecule 17β-estradiol were predicted by DAESTB. The table shows the top 1–25 ranked miRNAs associated with 17β-estradiol on the left and the bottom 26–50 ranked miRNAs on the right. As a result, 11(20) of the top 24(50) were confirmed by the recent experimental literatures in PubMed

miRNA	Evidence	miRNA	Evidence
hsa-mir-29c	PMID: 33070965	hsa-mir-181b-1	unconfirmed
hsa-mir-30a	PMID: 29331043	hsa-mir-99a	unconfirmed
hsa-mir-106b	PMID: 28422740	hsa-mir-34a	PMID: 33937961
hsa-mir-195	PMID: 26345235	hsa-mir-93	PMID: 23492819
hsa-mir-141	unconfirmed	hsa-mir-30d	PMID: 28066696
hsa-mir-125b-1	unconfirmed	hsa-mir-196a-1	unconfirmed
hsa-mir-16-1	unconfirmed	hsa-mir-101-1	unconfirmed
hsa-mir-16-2	unconfirmed	hsa-mir-302a	unconfirmed
hsa-mir-30e	PMID: 19223510	hsa-mir-301b	unconfirmed
hsa-mir-199a-1	unconfirmed	hsa-mir-331	unconfirmed
hsa-mir-20a	PMID: 21914226	hsa-mir-516b-2	unconfirmed
hsa-mir-145	PMID: 28011237	hsa-mir-34c	PMID: 24245576
hsa-mir-29b-1	PMID: 28466779	hsa-mir-29a	PMID: 22393047
hsa-mir-182	PMID: 34201250	hsa-mir-127	PMID: 34201250
hsa-mir-221	PMID: 30574741	hsa-mir-106a	unconfirmed
hsa-mir-210	PMID: 25503465	hsa-mir-329-2	unconfirmed
hsa-mir-199a-2	unconfirmed	hsa-mir-137	unconfirmed
hsa-mir-191	PMID: 29247596	hsa-mir-155	PMID: 23568502
hsa-mir-22	PMID: 24715036	hsa-mir-516a-2	unconfirmed
hsa-mir-129-2	unconfirmed	hsa-mir-125a	unconfirmed
hsa-mir-26b	PMID: 30690256	hsa-mir-150	unconfirmed
hsa-mir-373	unconfirmed	hsa-mir-100	PMID: 26908882
hsa-mir-24-1	unconfirmed	hsa-mir-214	PMID: 27884727
hsa-mir-494	PMID: 29867756	hsa-mir-199b	unconfirmed
hsa-mir-222	PMID: 27659519	hsa-mir-29b-2	unconfirmed

expression was much higher in LNCaP cells than in PC3 cells, and MiR-146a inhibitors were shown to inhibit apoptosis in 5-Aza-CdR-treated cells [63]. In WL-2 cells, miR-9-3 was fully methylated and 5-aza-2'-deoxycytidine treatment caused miR-9-3 promoter demethylation and pri-miR-9-3 re-expression [64].

The third small molecule is doxorubicin (DOX), which is a clinical first-line anti-cancer drug and is widely used by researchers [65]. After the DAESTB experiment, we obtained a ranking of all miRNAs associated with doxorubicin. 12 out of the top 20 miRNAs were validated, while 31 out of the top 50 miRNAs were validated in Table 3. For example, results from the study showed that miR-221 expression was upregulated after treatment of SCC4 and SCC9 cells with doxorubicin [66]. Using SUDHL9 (ABC-type) and OCI-Ly1 (GCB-type) cells, the effect of doxorubicin on reducing cell viability was enhanced by miR-197 transfection [67]. MiR-107 was identified as a target of circ-LTBP1, and by downregulating miR-107 abolished the circ-LTBP1-mediated response to doxorubicin-stimulated cells [68]. MiR-146a partially reversed doxorubicin-induced cardiotoxicity by targeting the TAF9b/P53 pathway to attenuate apoptosis and modulate autophagy levels [69].

Moreover, for the case study, we adopt an alternative approach to further demonstrate the predictive ability of DAESTB on isolated small molecule. We removed the known and verified small molecule–miRNA associations related to predictive small molecule. Namely, we set the row corresponding to the two small molecules 5-FU and DOX to 0 in turn in the known small molecule–miRNA associations matrix. It means that we only used similarity information and known small molecule–miRNA associations of the other small molecule to predict miRNA candidates. The predictive results are presented in supplementary material (Tables S4–S5). Some of the predictive miRNA candidates have been validated in SM2miR database and recent experimental literatures in Pubmeb. We also derived AUC curves for these small molecules as well as other assessment indicators as shown in supplementary material(Figure S3 and Table S6).

Discussion

The research has shown that miRNAs have an indispensable influence in biological and human diseases, as well as small molecule drugs that are crucial in the treatment of diseases. With the recent research, we can see that exploring the relationships between small molecule and miRNA will provide a guiding effect in disease diagnosis and treatment. In this work, we proposed an STB method based on deep autoencoder to predict the potential associations between small molecule–miRNA. First, we integrated small molecule–small molecule similarity, miRNA–miRNA similarity and known small molecule–miRNA associations to construct a high-dimensional feature representation matrix of combination pairs of small molecule–miRNA associations, and used the small molecule–miRNA associations as the corresponding labels of this high-dimensional matrix, with the aim at making the features of small molecule–miRNA properties more abundant. Second, the constructed high-dimensional feature attribute matrix was dimensionally reduced by a three-layer deep autoencoder to extract the potential feature representation between each combined small molecule–miRNA association pair, which aims to improve the efficiency of model learning as well as to remove noise. Finally, we used an STB method to predict potential associations between small molecule–miRNA. We used 5-fold CV and 10-fold CV for experimental comparison and validation on the dataset 1 and the dataset 2, respectively. Compared with other methods, the AUC values of DAESTB reached 0.9835 and 0.8637 on the dataset 1 and the dataset 2, respectively,

Table 2. Based on the dataset 1, the top 50 miRNAs associated with the small molecule 5-Aza-CdR were predicted by DAESTB. The table shows the top 1–25 ranked miRNAs associated with 5-Aza-CdR on the left and the bottom 26–50 ranked miRNAs on the right. As a result, 11(20) of the top 24(50) were confirmed by the recent experimental literatures in PubMed.

miRNA	Evidence	miRNA	Evidence
hsa-mir-18a	PMID: 23888958	hsa-mir-30c-2	unconfirmed
hsa-mir-222	PMID: 26975503	hsa-mir-30c-1	unconfirmed
hsa-mir-19b-1	unconfirmed	hsa-mir-30a	PMID: 26934121
hsa-mir-106a	PMID: 31115533	hsa-mir-148a	PMID: 21610744
hsa-mir-92a-1	unconfirmed	hsa-let-7a-3	unconfirmed
hsa-mir-143	PMID: 26351742	hsa-mir-146b	unconfirmed
hsa-mir-342	unconfirmed	hsa-mir-203a	unconfirmed
hsa-mir-10b	PMID: 21562367	hsa-mir-9-2	PMID: 25855800
hsa-mir-19b-2	unconfirmed	hsa-mir-23a	PMID: 25213664
hsa-mir-221	PMID: 23770133	hsa-mir-99b	unconfirmed
hsa-mir-186	PMID: 30793488	hsa-mir-100	unconfirmed
hsa-mir-181c	PMID: 20080834	hsa-mir-96	unconfirmed
hsa-mir-223	PMID: 27704586	hsa-mir-199a-2	unconfirmed
hsa-mir-150	unconfirmed	hsa-mir-30e	unconfirmed
hsa-mir-29c	PMID: 26975503	hsa-mir-142	PMID: 24236112
hsa-mir-199a-1	unconfirmed	hsa-mir-423	unconfirmed
hsa-mir-26b	PMID: 20806079	hsa-mir-144	unconfirmed
hsa-mir-101-1	unconfirmed	hsa-mir-204	PMID: 29850805
hsa-let-7d	PMID: 26802971	hsa-mir-206	unconfirmed
hsa-mir-146a	PMID: 24885368	hsa-mir-181b-2	unconfirmed
hsa-mir-199b	unconfirmed	hsa-mir-132	PMID: 22310291
hsa-mir-197	unconfirmed	hsa-mir-15b	PMID: 26934121
hsa-mir-181b-1	unconfirmed	hsa-mir-9-3	PMID: 25855800
hsa-mir-128-2	unconfirmed	hsa-mir-30d	unconfirmed
hsa-mir-224	PMID: 23770133	hsa-mir-133b	PMID: 26622593

Table 3. Based on the dataset 1, the top 50 miRNAs associated with the small molecule doxorubicin were predicted by DAESTB. The table shows the top 1–25 ranked miRNAs associated with doxorubicin on the left and the bottom 26–50 ranked miRNAs on the right. As a result, 12(20) of the top 31(50) were confirmed by the recent experimental literatures in PubMed.

miRNA	Evidence	miRNA	Evidence
hsa-let-7a-2	unconfirmed	hsa-mir-223	PMID: 31695022
hsa-mir-221	PMID: 28677788	hsa-mir-324	unconfirmed
hsa-let-7a-3	unconfirmed	hsa-mir-26b	PMID: 33767588
hsa-mir-125b-2	PMID: 19890372	hsa-mir-381	PMID: 30266665
hsa-let-7a-1	unconfirmed	hsa-mir-150	PMID: 31897096
hsa-mir-125b-1	PMID: 25451164	hsa-mir-30b	PMID: 35134991
hsa-mir-197	PMID: 29890998	hsa-mir-122	PMID: 27138141
hsa-mir-107	PMID: 35076816	hsa-mir-199b	PMID: 27033315
hsa-mir-373	unconfirmed	hsa-mir-132	PMID: 32505695
hsa-mir-146a	PMID: 31511497	hsa-mir-101-1	unconfirmed
hsa-mir-423	unconfirmed	hsa-mir-486	unconfirmed
hsa-mir-155	PMID: 26893712	hsa-mir-205	PMID: 35451580
hsa-mir-191	PMID: 20169152	hsa-mir-342	unconfirmed
hsa-let-7f-1	unconfirmed	hsa-mir-100	PMID: 27990096
hsa-mir-23a	PMID: 30831132	hsa-mir-146b	unconfirmed
hsa-mir-195	PMID: 31300525	hsa-mir-208a	PMID: 27990619
hsa-let-7d	PMID: 28665983	hsa-mir-99a	unconfirmed
hsa-mir-542	PMID: 30245930	hsa-mir-142	PMID: 32302291
hsa-mir-16-2	unconfirmed	hsa-mir-194-1	unconfirmed
hsa-mir-106b	PMID: 18212054	hsa-mir-15b	PMID: 28145098
hsa-let-7b	PMID: 25789066	hsa-mir-32	unconfirmed
hsa-mir-29c	unconfirmed	hsa-mir-18b	unconfirmed
hsa-let-7c	PMID: 33051247	hsa-mir-199a-2	unconfirmed
hsa-mir-125a	PMID: 27880721	hsa-mir-204	PMID: 33274007
hsa-mir-483	unconfirmed	hsa-mir-34a	PMID: 31235998

which were higher than those of other methods. The experimental results demonstrated that the proposed method DAESTB could better predict the potential associations between small molecule–miRNA.

Although DAESTB achieves good performance in predicting small molecule–miRNA associations, it still has some limitations. Firstly, few small molecule–miRNA associations are collected in existing databases, and data on the side effects of small molecules and diseases from the current databases are incomplete; predictions based on these incomplete annotations will miss or overlook some small molecule–miRNA associations. In future research, the increase in more newly identified small molecule-related side effects and diseases information being collected will advances the development of computational approaches to predict small molecule–miRNA associations. Secondly, there is still room for improvement for data collection, if the miRNA sequence similarity and affiliation information of small molecules could be considered to construct more complete data, the prediction results would be more accurately. Thirdly, if the associations between small molecule and miRNA can be used to further predict the associations between miRNA and disease, to achieve an effective integration of small molecule drugs to miRNAs and then to diseases, it will be possible to further improve the precision of drug-targeted therapy.

Key Points

- As miRNAs targeting small molecules are closely associated with the progression of various human diseases, it is critical to develop effective computational predictors to predict the relevance of small molecules to miRNAs.
- We propose a new prediction method, DAESTB, which improves prediction efficiency by removing noise through deep autoencoder dimensionality reduction and identifies potential associations from new small molecule–miRNA associations networks using a scalable tree boosting methods.
- The performance on widely used datasets shows that DAESTB outperforms other advanced prediction methods. In addition, several case studies have shown that DAESTB is also able to accurately identify novel small molecule–miRNA associations.

Data availability

The data and source code can be freely downloaded from: https://github.com/biohnuster/DAESTB.

Acknowledgments

We would like to thank the anonymous reviewers for their valuable comments and improvement suggestions that further lead us to improve this paper.

Funding

National Natural Science Foundation of China (Grant No.61902125, 62002111,62202162); Scientific Research Project of Hunan Education Department (Grant No.19C1788, 21B0453).

References

1. Bartel DP. Micrornas: genomics, biogenesis, mechanism, and function. *Cell* 2004;**116**(2):281–97.
2. Beermann J, Piccoli M-T, Viereck J, *et al.* Non-coding rnas in development and disease: background, mechanisms, and therapeutic approaches. *Physiol Rev* 2016;**96**(4):1297–1325.
3. Chen X, Xie D, Zhao Q, *et al.* Micrornas and complex diseases: from experimental results to computational models. *Brief Bioinform* 2019;**20**(2):515–39.
4. Zhang S, Chen L, Jung EJ, *et al.* Targeting micrornas with small molecules: from dream to reality. *Clinical Pharmacology & Therapeutics* 2010;**87**(6):754–8.
5. Chen X, Wang L, Jia Q, *et al.* Predicting mirna–disease association based on inductive matrix completion. *Bioinformatics* 2018;**34**(24):4256–65.
6. Chen X, Yin J, Jia Q, *et al.* Mdhgi: matrix decomposition and heterogeneous graph inference for mirna-disease association prediction. *PLoS Comput Biol* 2018;**14**(8):e1006418.
7. Chen X, Xie D, Wang L, *et al.* Bnpmda: bipartite network projection for mirna–disease association prediction. *Bioinformatics* 2018;**34**(18):3178–86.
8. Huang L, Zhang L, Chen X. Updated review of advances in micrornas and complex diseases: towards systematic evaluation of computational models. *Brief Bioinform* 2022:1–15.
9. Peng L, Yang C, Huang L, *et al.* Rnmflp: Predicting circrna–disease associations based on robust nonnegative matrix factorization and label propagation. *Brief Bioinform* 2022;**23**(5):1–14.
10. Chen J, Peng H, Han G, *et al.* Hogmmnc: a higher order graph matching with multiple network constraints model for gene–drug regulatory modules identification. *Bioinformatics* 2019;**35**(4):602–10.
11. Chen X, Sun Y-Z, Zhang D-H, *et al.* Nrdtd: a database for clinically or experimentally supported non-coding rnas and drug targets associations. *Database* 2017;**2017**:1–6.
12. Chen X, Guan N-N, Sun Y-Z, *et al.* Microrna-small molecule association identification: from experimental results to computational models. *Brief Bioinform* 2020;**21**(1):47–61.
13. Wang C-C, Chen X. A unified framework for the prediction of small molecule–microrna association based on cross-layer dependency inference on multilayered networks. *J Chem Inf Model* 2019;**59**(12):5281–93.
14. Yin J, Chen X, Wang C-C, *et al.* Prediction of small molecule–microrna associations by sparse learning and heterogeneous graph inference. *Mol Pharm* 2019;**16**(7):3157–66.
15. Lv Y, Wang S, Meng F, *et al.* Identifying novel associations between small molecules and mirnas based on integrated molecular networks. *Bioinformatics* 2015;**31**(22):3638–44.
16. Li J, Lei K, Zengrui W, *et al.* Network-based identification of micrornas as potential pharmacogenomic biomarkers for anticancer drugs. *Oncotarget* 2016;**7**(29):45584.
17. Jia Q, Chen X, Sun Y-Z, *et al.* Inferring potential small molecule–mirna association based on triple layer heterogeneous network. *J Chem* 2018;**10**(1):30.
18. 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.
19. Jia Q, Chen X, Sun Y-Z, *et al.* In silico prediction of small molecule-mirna associations based on the hetesim algorithm. *Molecular Therapy-Nucleic Acids* 2019;**14**:274–86.
20. Wang C-C, Chen X, Jia Q, *et al.* Rfsmma: a new computational model to identify and prioritize potential small molecule–mirna associations. *J Chem Inf Model* 2019;**59**(4):1668–79.
21. Zhao Y, Chen X, Yin J, *et al.* Snmfsmma: using symmetric nonnegative matrix factorization and kronecker regularized least squares to predict potential small molecule-microrna association. *RNA Biol* 2020;**17**(2):281–91.
22. Chen X, Zhou C, Wang C-C, *et al.* Predicting potential small molecule–mirna associations based on bounded nuclear norm regularization. *Brief Bioinform* 2021;**22**(6):1–14.

23. Wang C-C, Zhu C-C, Chen X. Ensemble of kernel ridge regression-based small molecule–mirna association prediction in human disease. *Brief Bioinform* 2022;**23**(1):1–11.
24. Wang S-H, Wang C-C, Huang L, *et al*. Dual-network collaborative matrix factorization for predicting small molecule-mirna associations. *Brief Bioinform* 2022;**23**(1):1–12.
25. Liu X, Wang S, Meng F, *et al*. Sm2mir: a database of the experimentally validated small molecules' effects on microrna expression. *Bioinformatics* 2013;**29**(3):409–11.
26. Knox C, Law V, Jewison T, *et al*. Drugbank 3.0: a comprehensive resource for 'omics' research on drugs. *Nucleic Acids Res* 2010;**39**(suppl_1):D1035–41.
27. Wang Y, Lee CGL. Microrna and cancer–focus on apoptosis. *J Cell Mol Med* 2009;**13**(1):12–23.
28. Ming L, Zhang Q, Deng M, *et al*. An analysis of human microrna and disease associations. *PloS one* 2008;**3**(10):e3420.
29. Jiang Q, Wang Y, Hao Y, *et al*. mir2disease: a manually curated database for microrna deregulation in human disease. *Nucleic Acids Res* 2009;**37**(suppl_1):D98–104.
30. Ruepp A, Kowarsch A, Schmidl D, *et al*. Phenomir: a knowledgebase for microrna expression in diseases and biological processes. *Genome Biol* 2010;**11**(1):1–11.
31. Gottlieb A, Stein GY, Ruppin E, *et al*. Predict: a method for inferring novel drug indications with application to personalized medicine. *Mol Syst Biol* 2011;**7**(1):496.
32. Lv S, Li Y, Wang Q, *et al*. A novel method to quantify gene set functional association based on gene ontology. *Journal of The Royal Society Interface* 2012;**9**(70):1063–72.
33. Hattori M, Okuno Y, Goto S, *et al*. Development of a chemical structure comparison method for integrated analysis of chemical and genomic information in the metabolic pathways. *J Am Chem Soc* 2003;**125**(39):11853–65.
34. Köhler S, Bauer S, Horn D, *et al*. Walking the interactome for prioritization of candidate disease genes. *The American Journal of Human Genetics* 2008;**82**(4):949–58.
35. Zhu F, Shi Z, Qin C, *et al*. Therapeutic target database update 2012: a resource for facilitating target-oriented drug discovery. *Nucleic Acids Res* 2012;**40**(D1):D1128–36.
36. Kuhn M, Campillos M, Letunic I, *et al*. A side effect resource to capture phenotypic effects of drugs. *Mol Syst Biol* 2010;**6**(1): 343.
37. Deepthi K, Jereesh AS. An ensemble approach for circrna-disease association prediction based on autoencoder and deep neural network. *Gene* 2020;**762**:145040.
38. Chen T, Guestrin C. Xgboost: A scalable tree boosting system. In: *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. ACM Press, New York, USA, 2016, 785–94.
39. Liu D, Huang Y, Nie W, *et al*. Smalf: mirna-disease associations prediction based on stacked autoencoder and xgboost. *BMC bioinformatics* 2021;**22**(1):1–18.
40. Hervé A, Lynne JW. Principal component annalysis. *Wiley interdisciplinary reviews computational statistic* 2010;**2**(4):433–59.
41. Van der Maaten L, Hinton G. Visualizing data using t-sne. *Journal of machine learning research* 2008;**9**(11):2579–2605.
42. McInnes L, Healy J, Melville J. *Umap: Uniform manifold approximation and projection for dimension reduction*arXiv preprint arXiv:1802.03426, 2018.
43. Brown TA. In: David A. Kenny (ed), *Confirmatory factor analysis for applied research*. Guilford publications, New York, 2015.
44. Stone JV. Independent component analysis: an introduction. *Trends Cogn Sci* 2002;**6**(2):59–64.
45. Wang G, Wang R. Sparse coding network model based on fast independent component analysis. *Neural Computing and Applications* 2019;**31**(3):887–93.
46. Maher M. Logistic regression in data analysis: an overview. *International Journal of Data Analysis Techniques and Strategies* 2011;**3**(3):281–99.
47. Leung KM. Naive bayesian classifier. *Polytechnic University Department of Computer Science/Finance and Risk Engineering* 2007;**2007**:123–56.
48. Chen X, Zhu C-C, Yin J. Ensemble of decision tree reveals potential mirna-disease associations. *PLoS Comput Biol* 2019;**15**(7):e1007209.
49. Freund Y, Schapire RE. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences* 1997;**55**(1):119–39.
50. Breiman L. Random forests. *Machine learning* 2001;**45**(1):5–32.
51. Natekin A, Knoll A. Gradient boosting machines, a tutorial. *Front Neurorobot* 2013;**7**:21.
52. Hearst MA, Dumais ST, Osuna E, *et al*. Support vector machines. *IEEE Intelligent Systems and their applications* 1998;**13**(4):18–28.
53. Ke G, Meng Q, Finley T, *et al*. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems* 2017;**30**:3149–57.
54. Prokhorenkova L, Gusev G, Vorobev A, *et al*. Catboost: unbiased boosting with categorical features. In: *Advances in neural information processing processing systems*. Curran Associations Inc, New York, USA, 2018, p. 6638–48.
55. Simpson E, Santen RJ. Celebrating 75 years of oestradiol. *J Mol Endocrinol* 2015;**55**(3):T1–20.
56. Chuang T-D, Rehan A, Khorram O. Functional role of the long noncoding rna x-inactive specific transcript in leiomyoma pathogenesis. *Fertil Steril* 2021;**115**(1):238–47.
57. Koganti PP, Wang J, Cleveland B, *et al*. Estradiol regulates expression of mirnas associated with myogenesis in rainbow trout. *Mol Cell Endocrinol* 2017;**443**:1–14.
58. Zhang C, Zhao J, Deng H. 17β-estradiol up-regulates mir-155 expression and reduces tp53inp1 expression in mcf-7 breast cancer cells. *Mol Cell Biochem* 2013;**379**(1):201–11.
59. Ocadiz-Delgado R, Cruz-Colin J-L, Alvarez-Rios E, *et al*. Expression of mir-34a and mir-15b during the progression of cervical cancer in a murine model expressing the hpv16 e7 oncoprotein. *J Physiol Biochem* 2021;**77**(4):547–55.
60. Patra SK, Bettuzzi S. Epigenetic dna-(cytosine-5-carbon) modifications: 5-aza-2′-deoxycytidine and dna-demethylation. *Biochemistry (Moscow)* 2009;**74**(6):613–9.
61. Li D, Deng T, Li H, *et al*. Mir-143 and mir-135 inhibitors treatment induces skeletal myogenic differentiation of human adult dental pulp stem cells. *Arch Oral Biol* 2015;**60**(11):1613–7.
62. Kim K, Lee H-C, Park J-L, *et al*. Epigenetic regulation of microrna-10b and targeting of oncogenic mapre1 in gastric cancer. *Epigenetics* 2011;**6**(6):740–51.
63. Wang X, Gao H, Ren L, *et al*. Demethylation of the mir-146a promoter by 5-aza-2′-deoxycytidine correlates with delayed progression of castration-resistant prostate cancer. *BMC Cancer* 2014;**14**(1):1–11.
64. Qi Zhang L, Wang Q, Wong KY, *et al*. Infrequent dna methylation of mir-9-1 and mir-9-3 in multiple myeloma. *J Clin Pathol* 2015;**68**(7):557–61.
65. Zhu L, Lin M. The synthesis of nano-doxorubicin and its anticancer effect. *Anti-Cancer Agents in Medicinal Chemistry (Formerly Current Medicinal Chemistry-Anti-Cancer Agents)* 2021;**21**(18): 2466–77.

66. Liangzhi D, Ma S, Wen X, *et al.* Oral squamous cell carcinoma cells are resistant to doxorubicin through upregulation of mir-221. *Mol Med Rep* 2017;**16**(3): 2659–67.
67. Yang JM, Jang J-Y, Jeon YK, *et al.* Clinicopathologic implication of microrna-197 in diffuse large b cell lymphoma. *J Transl Med* 2018;**16**(1):1–14.
68. Li C, Zhang L, Xingpeng B, *et al.* Circ-ltbp1 is involved in doxorubicin-induced intracellular toxicity in cardiomyocytes via mir-107/adcy1 signal. *Mol Cell Biochem* 2022;**477**(4):1127–38.
69. Pan J-A, Tang Y, Jian-Ying Y, *et al.* mir-146a attenuates apoptosis and modulates autophagy by targeting taf9b/p53 pathway in doxorubicin-induced cardiotoxicity. *Cell Death Dis* 2019;**10**(9): 1–15.