frontiers | Frontiers in Genetics

TYPE Original Research
PUBLISHED 02 December 2022
DOI 10.3389/fgene.2022.1079053

OPEN ACCESS

EDITED BY
Leyi Wei,
Shandong University, China

REVIEWED BY
Yongqiang Xing,
Inner Mongolia University of Science and Technology, China
Fei Guo,
Central South University, China

*CORRESPONDENCE
Jianwei Li,
lijianwei@hebut.edu.cn

SPECIALTY SECTION
This article was submitted to Computational Genomics, a section of the journal Frontiers in Genetics

RECEIVED 25 October 2022
ACCEPTED 21 November 2022
PUBLISHED 02 December 2022

CITATION
Li J, Lin H, Wang Y, Li Z and Wu B (2022), Prediction of potential small molecule–miRNA associations based on heterogeneous network representation learning.
Front. Genet. 13:1079053.
doi: 10.3389/fgene.2022.1079053

Prediction of potential small molecule–miRNA associations based on heterogeneous network representation learning

Jianwei Li[1,2]*, Hongxin Lin[1], Yinfei Wang[1], Zhiguang Li[1] and Baoqin Wu[1]

[1]School of Artificial Intelligence, Institute of Computational Medicine, Hebei University of Technology, Tianjin, China, [2]Hebei Province Key Laboratory of Big Data Calculation, Hebei University of Technology, Tianjin, China

MicroRNAs (miRNAs) are closely associated with the occurrences and developments of many complex human diseases. Increasing studies have shown that miRNAs emerge as new therapeutic targets of small molecule (SM) drugs. Since traditional experiment methods are expensive and time consuming, it is particularly crucial to find efficient computational approaches to predict potential small molecule-miRNA (SM-miRNA) associations. Considering that integrating multi-source heterogeneous information related with SM-miRNA association prediction would provide a comprehensive insight into the features of both SMs and miRNAs, we proposed a novel model of Small Molecule-MiRNA Association prediction based on Heterogeneous Network Representation Learning (SMMA-HNRL) for more precisely predicting the potential SM-miRNA associations. In SMMA-HNRL, a novel heterogeneous information network was constructed with SM nodes, miRNA nodes and disease nodes. To access and utilize of the topological information of the heterogeneous information network, feature vectors of SM and miRNA nodes were obtained by two different heterogeneous network representation learning algorithms (HeGAN and HIN2Vec) respectively and merged with connect operation. Finally, LightGBM was chosen as the classifier of SMMA-HNRL for predicting potential SM-miRNA associations. The 10-fold cross validations were conducted to evaluate the prediction performance of SMMA-HNRL, it achieved an area under of ROC curve of 0.9875, which was superior to other three state-of-the-art models. With two independent validation datasets, the test experiment results revealed the robustness of our model. Moreover, three case studies were performed. As a result, 35, 37, and 22 miRNAs among the top 50 predicting miRNAs associated with 5-FU, cisplatin, and imatinib were validated by experimental literature works respectively, which confirmed the effectiveness of SMMA-HNRL. The source code and experimental data of SMMA-HNRL are available at https://github.com/SMMA-HNRL/SMMA-HNRL.

KEYWORDS

small molecule-miRNA association prediction, heterogeneous information, heterogeneous network representation learning, machine learning, lightgbm

Introduction

MicroRNAs (miRNAs) are a large group of non-coding RNAs (ncRNAs) with approximately 22 nucleotides in length, which are widespread in eukaryotes (Bartel, 2004). Since the discovery of the first miRNA, lin-4, in Caenorhabditis elegans by Lee et al. (1993), accumulating evidence has demonstrated that miRNAs play vital roles in various key physiological processes, including cell proliferation (Cheng et al., 2005), cell differentiation (Miska, 2005), cell apoptosis (Xu et al., 2004), regulation of animal immune function (Stern-Ginossar et al., 2007) and regulation of gene expression levels (Shivdasani, 2006) etc. Meanwhile, many studies have confirmed that numerous complex human diseases are also closely related to the dysregulations of related key miRNAs (Croce and Calin, 2005; Sayed and Abdellatif, 2011; Chen et al., 2019). Due to the ubiquity of miRNAs in physiological and pathological processes, miRNAs are also recognized as a potentially important class of drug targets (Liu et al., 2008; Rossi, 2009; Cheng et al., 2015). Nowadays, computer-aided drug design has been applied broadly in the early stages of drug development, and the prediction results of computational models can provide directions for researchers to find the most effective drugs, reduce experimental costs and blindness significantly (Zhao et al., 2019). Among them, computational prediction of Small Molecule-miRNA (SM-miRNA) associations is a critical step in drug R&D (Chen et al., 2018). With deepening research in the field of SM-miRNA association prediction, many corresponding databases have been constructed, such as SM2miR (Liu et al., 2013), NoncoRNA (Li et al., 2020), mTD (Chen et al., 2017a), and NRDTD (Chen et al., 2017b). These databases provide abundant resources for exploring SM-miRNA associations and make it possible to construct effective and accurate SM-miRNA association prediction models. SM-miRNA computational models are usually divided into three categories, models based on biological networks, models based on machine learning algorithms and other prediction models.

In the first category, the models construct biological networks based on biological information and utilize network topology information to predict potential SM-miRNA associations. In 2015, Lv et al. (2015) built an integrated heterogeneous network by SM and miRNA similarity networks and SM-miRNA association network. They employed random walk with restart (RWR) algorithms on the heterogeneous network for predicting potential SM-miRNA associations. In 2016, Li et al. (2016) developed a network-based inference framework which was termed SMiR-NBI. A heterogeneous network which consisted of SMs, miRNAs and genes was constructed, and the network-based inference algorithm was implemented to calculate the association scores between the given SMs and miRNAs. Qu et al. (2018) proposed a prediction framework based on a heterogeneous network which was named TLHNSMMA in 2018. TLHNSMMA constructed a triple-layer network and finally predicted potential SM-miRNA associations by the iterative update algorithm based on the global network. In the same year, Guan et al. (2018) proposed GISMMA based on Graphlet interactions. Graphlet interactions between SMs and miRNAs were calculated on SM and miRNA similarity networks. In 2020, Shen et al. (2020) proposed a computational model named SMMART based on graph regularization technique.

The second category of the computational models predicts novel SM-miRNA associations based on machine learning algorithms. Extracting the biological features of SMs and miRNAs for training the classifiers, potential associations are predicted with machine learning algorithms. In 2019, Wang et al. (2019) proposed RFSMMA model based on random forest algorithm. A filtering approach was employed to extract reliable features of SM-miRNA pairs by using their similarity data. Subsequently, the features were exploited to train the random forest model, and potential SM-miRNA associations were predicted with it. In 2022, Wang et al. (2022) developed an EKRRSMMA model based on ensemble of kernel ridge regression. By constructing different feature subsets for SMs and miRNAs, an integrated learning model containing multiple KRR-based base learning tasks was constructed. The prediction results of all base learners were averaged and the results were introduced as the SM-miRNA association scores. Beside the above two categories, there are also other models which can predict SM-miRNA associations. In 2019, Xie et al. (2019) proposed a new text mining framework, termed EmDL, for extracting associations between miRNAs and SMs efficacy from the literature and recording them in the database. In 2012, Jiang et al. (2012) constructed a SM-miRNA Network (SMirN) for each type of 23 common cancers. The associations of cancer-related miRNAs with SMs were determined by the enrichment scores. To give readers a clear overview, Supplementary Material S1 summarizes the aforementioned models in a tabular form.

More recently, network representation learning algorithms have been widely used in the field of biomedical sciences (Yue et al., 2020). In 2021, Thafar et al. (2021) predicted drug-target associations with node2vec (Grover and Leskovec, 2016) and ensemble learning. Ji et al. (2020) adopted the LINE (Tang et al., 2015) to catch the feature information from the drug-target network and utilized random forest method as the classifier in 2020. Early network representation learning algorithms could only address homogeneous networks. Yet in the reality, a vast number of networks are composed of different types of entities

and different kinds of relationships, which were called heterogeneous information networks. Heterogeneous network representation learning algorithm was more capable of retaining the rich structural and semantic information in heterogeneous information networks. Thus, numerous heterogeneous network representation learning algorithms have been proposed rapidly, and have been implemented into biological networks. In 2021, Deng et al. (2021) used HIN2Vec (Fu et al., 2017) to learn the embedding vectors for each node in lncRNA-disease-miRNA heterogeneous network and utilized gradient boosting tree (GBT) classifier for predicting potential lncRNA-disease associations. In 2018, Zhu et al. (2018) utilized Metapath2vec (Dong et al., 2017) to extract heterogeneous network features and employ a kernelized Bayesian matrix factorization method for predicting drug-gene associations.

In this study, we proposed a novel model, SM-MiRNA Association prediction based on Heterogeneous Network Representation Learning (SMMA-HNRL), to improve the performance of SM-miRNA association prediction. The data was collected from six networks (miRNA-SM, miRNA-disease, miRNA-miRNA, SM-disease, SM-SM, disease-disease) for constructing miRNA-SM-disease heterogeneous information network. Inspired by the success of integrated features on the lncRNA-disease association prediction problem (Li et al., 2021), we employed two excellent heterogeneous network representation learning algorithms, HIN2Vec (Fu et al., 2017) and HeGAN (Hu et al., 2019), to embed all nodes of the miRNA-SM-disease heterogeneous network into low-dimensional vectors respectively, and then combined them into the novel feature vectors of SMs and miRNAs. Finally, Hadamard function was chosen to gain all SM-miRNA vector pairs, and LightGBM (Ke et al., 2017) classifier was selected to predict potential SM-miRNA associations. To assess the prediction performance of SMMA-HNRL, we compared it with three state-of-the-art models with 10-fold-cross validations. For validating the robustness, our model performed on two independent validation datasets. Moreover, the dependable prediction performance of SMMA-HNRL was also confirmed with three case studies. All the results of evaluation experiments demonstrated the reliable and predictive performance of SMMA-HNRL.

Materials and methods

SM-miRNA association network

The experimentally validated SM-miRNA associations used in our study was downloaded from the SM2miR v3.0 database (Liu et al., 2013). By manual inspection, we eliminated the SMs which were not present in DrugBank (Wishart et al., 2018), and merged the mature miRNAs which were generated from the same precursor miRNAs (e.g., hsa-miR-21-3p and hsa-miR-21-5p). Then, the format of mature miRNAs was converted to that of precursor miRNA. Moreover, non-human data and duplicate SM-miRNA associations were culled out. Finally, 1766 experimentally validated SM-miRNA associations which included 546 miRNAs and 93 SMs were obtained. Finally, an SM-miRNA association network was constructed based on these 1766 associations which was used during training the model and the cross-validation evaluation.

miRNA-disease association network

Human experimentally validated miRNA-disease associations was downloaded from the HMDD v3.2 database (Huang et al., 2019), and disease names of miRNA-disease associations were converted into the standardized names according to the MESH glossary. After removing duplicated data, a total of 18,732 miRNA-disease associations involving 1206 miRNAs and 892 diseases were obtained and the miRNA-disease association network was constructed with them.

SM-disease association network

The SM-disease association data was collected from the SCMFDD-L dataset in the SCMFDD database (Zhang et al., 2018). SCMFDD acquired available drug-disease associations from the CTD database (Davis et al., 2017) and selected drugs with known drug substructure information. The SM drugs were selected and duplicated data was removed. Through screening, 49,032 pairs of SM-disease associations were obtained which included 1313 SMs and 2822 diseases.

Integrated SM similarity network

The integrated SM similarity data was downloaded from the DrugSimDB database (Azad et al., 2021) which were the mean values of chemical structure similarity, target protein sequence-based similarity, target protein functional similarity and drug-induced pathway similarity of SM drugs. The integrated SM similarity network was constructed according to this data which included 1331 SM drugs.

Integrated miRNA similarity network

The miRNA similarity data was sourced from miRNA-disease associations and Gene Ontology (GO) annotations of miRNA target genes respectively. In 2010, Wang et al. (2010) proposed a method named MISIM to calculate the functional similarity of miRNAs based on the hypothesis that functionally similar miRNAs were often associated with semantically similar diseases. In 2019, MISIM was

updated and named MISIM v2.0 by our research group which not only had a threefold increase in data content compared with MISIM but also improved the original MISIM algorithm (Li et al., 2019). Additionally, Yang et al. (2018) developed a novel method called MIRGOFS which calculated the functional similarity of miRNAs based on the GO annotations of their target genes. We downloaded miRNA similarity data from MISIM v2.0 and the normalized miRNA similarity network data from MIRGOFS respectively. To facilitate calculation, mature miRNAs produced from the same pre-miRNA were merged and converted into the precursor miRNA.

Inspired by SM similarity network, we integrated the above two miRNA similarity networks with average ensemble method. If one miRNA was in only one similarity network, its similarity value was considered the final results. In the end, an integrated miRNA similarity network consisting of 1309 miRNAs was obtained.

Disease semantic similarity network

The semantic similarity values between two diseases can be calculated based on the Medical Subject Headings (MESH) disease structure. Each disease represented by the MESH descriptor, which were obtained from National Library of Medicine (https://www.nlm.nih.gov/), could be represented as a Directed Acyclic Graph (DAG). One disease d can be denoted in the DAG as follows:

$$DAG_d = (d, T_d, E_d) \tag{1}$$

Where T_d represented the node set which was composed of disease d and all its ancestor nodes; and E_d represented the edge set of disease d in the DAG. The semantic contribution value D of disease t to disease d can be defined by the following equation:

$$\begin{cases} D_d(t) = 1 \\ D_d(t) = max\{\Delta * D_d(t') \mid t' \in \text{children of } t\} \text{ if } t \neq d \end{cases} \tag{2}$$

Eq. (2) indicated that if there were multiple paths for disease t to reach disease d in the DAG graph, the shortest path needed to be selected to achieve the maximum semantic contribution value. Δ was the semantic contribution factor which reflected the influence degree of the parent node on the child nodes in the DAG graph. Based on the related study by Xuan et al. (2013), the value of Δ was set as 0.5 in the beginning of the calculation. After accumulating the semantic contribution values of all disease nodes in the DAG, the semantic contribution value of every disease was obtained.

$$DV(d) = \sum_{t \in T_d} D_d(t) \tag{3}$$

With the semantic contribution value of each disease, we could calculate the similarity between any two diseases d_i and d_j according to Eq. (4).

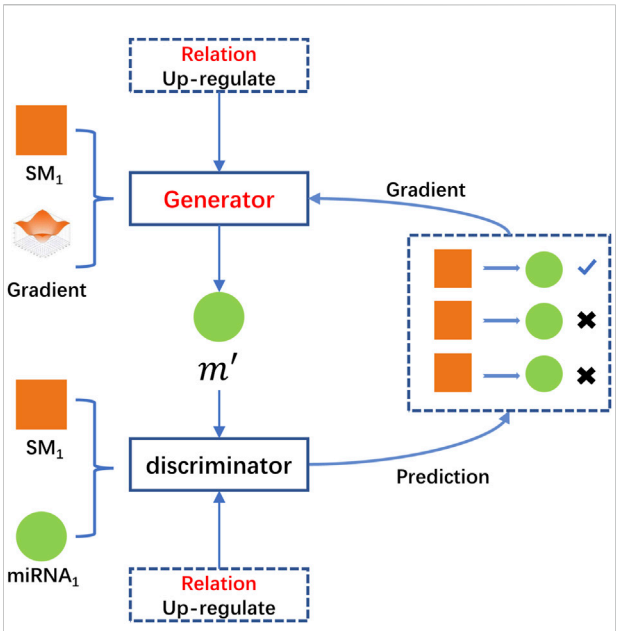

FIGURE 1
The framework of the HeGAN algorithm.

$$\text{Sim}(d_i, d_j) = \frac{\sum_{t \in T_{d_i} \cap T_{d_j}} \left(D_{d_i}(t) + D_{d_j}(t)\right)}{\text{DV}(d_i) + \text{DV}(d_j)} \tag{4}$$

where t was used to denote nodes of disease d_i and d_j in the DAG structure, $\text{DV}(d_i)$ and $\text{DV}(d_j)$ represented semantic values of disease d_i and d_j, and $D_{d_i}(t)$ and $D_{d_j}(t)$ were indicated as the semantic contributions of disease t to diseases d_i and d_j.

Feature extraction by HeGAN

HeGAN was the first method which introduced Generative Adversarial Networks (GAN) into heterogeneous networks representation learning problem (Hu et al., 2019). The basic idea of GAN was to train the discriminator and the generator with the ideas of competition, and thus obtained the data latent distribution. Compared with traditional heterogeneous network representation learning methods, HeGAN exhibited more stability to sparse data or noisy data and achieved the best performance in downstream tasks on public datasets. Besides, it should be noted that HeGAN did not employ meta-paths and there was no costly meta-path setup.

HeGAN was composed of two main competing modules, the relational perception discriminator and the generalized generator. For a given node, the generalized generator firstly attempted to generate fake samples associated with the given node and fed these fake samples to the discriminator. The

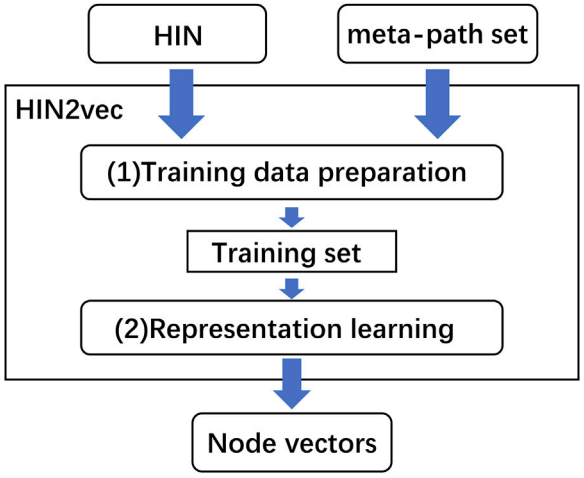

FIGURE 2
The flowchart of HIN2Vec algorithm.

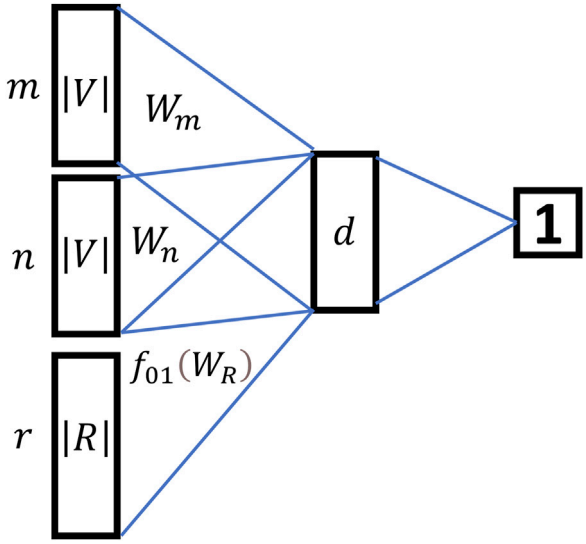

FIGURE 3
The neural network structure of HIN2Vec.

discriminator accepted true samples from the real network and fake samples generated by the generator, respectively. Secondly, HeGAN attempted to adjust its parameters to separate the fake samples from the true samples repeatedly. Finally, the discriminator predicted the probability of two nodes that there was a relationship r between them. In the iterative process, the trained discriminator continually forced the generator to generate better fake samples, while the discriminator would enhance its judgment ability correspondingly. Figure 1. illustrates the framework of the HeGAN algorithm.

Traditional network representation learning methods were limited performance due to lack of making full use of the valuable semantic information of heterogeneous information networks. For a given miRNA m, suppose that there were two nodes SM_1 and SM_2 which were associated with it. The traditional methods simply regarded SM_1 and SM_2 as true nodes and did not analyze them in depth. Normally, SM_1 and SM_2 were generally associated with m due to multiple reasons, such as SM_1 upregulated m, while SM_2 downregulated m. The traditional methods did not take full use of the valuable semantics embedded in heterogeneous networks, which would lower the accuracy of functional predictions of SM_1 and SM_2. The relational perception discriminator and generalized generator introduced by HeGAN were more suitable for distinguishing various types of semantic relations between two nodes. Besides that, the negative samples of the traditional methods were limited to the number of known samples. In practice, the most representative negative samples were likely to be located between the embedding vectors corresponding to existing nodes, not the existing nodes. To better generate negative samples, HeGAN specifically introduced one generalized generator to generate negative nodes which did not exist in the samples. For example, m' in Figure 1, it did not exist in the original graph, but it was the node which best represented the original network.

Feature extraction by HIN2Vec

HIN2Vec was another heterogeneous network representation learning method with excellent performance (Fu et al., 2017). The core part of HIN2Vec was one three-layer neural network which learned the rich information from the heterogeneous information network by captured different relationship information of network topologies. HIN2Vec not only obtained low-dimensional representations of nodes, but also learned representations of relationships (meta-paths) in the networks. HIN2Vec also got the best performance in downstream tasks. The flowchart of the HIN2Vec algorithm is shown in Figure 2.

As illustrated in Figure 2, the HIN2Vec model consisted of two main parts, one was the training data preparation which was generated based on random walk and negative sampling, the other was representation learning which was performed on the generated training data. In training data generation, HIN2Vec represented the heterogeneous network in the form of $\langle m, n, r, L(m, n, r)\rangle$, where m, n represented two nodes, r was a different type of relationship between two nodes, $L(m, n, r)$ was a binary value representing whether there was a relationship between the m and n nodes. HIN2Vec utilized random walk algorithm to generate node sequences and differentiated their types r. It was unlike Metapath2vec (Dong et al., 2017) which

walked exactly according to a given meta-path, the HIN2Vec model completely randomly selected different walking nodes. If there was a connection between two nodes, a random walk could be conducted. Considering the above, HIN2Vec would retain more contextual information and acquire richer semantics.

In the representation learning part, HIN2Vec innovatively transformed the relationship between two nodes from multi-classification problem to a multiple binary classification problem. HIN2Vec built a three-layer feedforward neural network as a logical binary classifier to predict whether there is a definite relationship r between two nodes which avoided traversing all relationships in the network, and learned the vector representation of nodes and relationships at the same time. Figure 3 showed the neural network structure of HIN2Vec.

From Figure 3, the relationship between two nodes in HIN2Vec was no longer considered as a prediction object, but as training data of the input layer. The model mainly predicted whether there was a specific relationship r between node m and node n.The inputs to the model were three one-hot vectors, $\vec{m}, \vec{n}$ and $\vec{r}$. They were converted in the latent layer to the latent vector $W_M^{'}\vec{m}, W_N^{'}\vec{n}$ and $f_{01}(W_R^{'}\vec{r})$. Since the semantic information of the node was different from the semantic information of the relationship, the regularization function $f_{01}(.)$ was added before the relationship r for regularization to ensure that the value of the relationship r was between 0 and 1. Then the three latent vectors were aggregated with the Hadamard function (the elements in the vector were multiplied two by two) to obtain the form of $W_M^{'}\vec{m} \odot W_N^{'}\vec{n} \odot f_{01}(W_R^{'}\vec{r})$, and applied the identity function to activate. At the output layer, HIN2Vec took summation for the output d-dimensional vectors in the hidden layer and activated them with the Sigmoid function. Eventually, sigmoid $(\sum W_M^{'}\vec{m} \odot W_N^{'}\vec{n} \odot f_{01}(W_R^{'}\vec{r}))$ was utilized for logical classification.

HIN2Vec was trained iteratively on the training set D with a backpropagation algorithm with stochastic gradient descent. By continuously adjusting the weights of each entry W_m, W_n and W_R in set D, the objective function O was maximized, which was the multiplication of each training data entity $O_{m,n,r}(m,n,r)$ in set D. To simplify computation, HIN2Vec maximizes $\log O$ instead of directly maximizing O. The objective functions O and $\log O$ were defined as follows:

$$O \propto \log O = \sum_{m,n,r \in D} \log O_{m,n,r}(m,n,r) \tag{5}$$

In particular, in a training sample $\langle m,n,r,L(m,n,r)\rangle$, if $L(m,n,r)$ was 1, $O_{m,n,r}(m,n,r)$ aimed to maximize $P(r \mid m,n)$. Otherwise, the $O_{m,n,r}(m,n,r)$ aimed to minimize $P(r \mid m,n)$. $O_{m,n,r}(m,n,r)$, $\log O_{m,n,r}(m,n,r)$ and $P(r \mid m,n)$ were derived by the following formula:

$$O_{m,n,r}(m,n,r) = \begin{cases} P(r \mid m,n), \text{ if } L(m,n,r) = 1 \\ 1 - P(r \mid m,n), \text{ if } L(m,n,r) = 0 \end{cases} \tag{6}$$

$$\log O_{m,n,r}(m,n,r) = L(m,n,r)\log P(r \mid m,n) + [1 - L(m,n,r)]\log[1 - P(r|m,n)] \tag{7}$$

$$P(r|m,n) = \text{sigmoid}\left(\sum W_m^{'}\vec{m} \odot W_n^{'}\vec{n} \odot f_{01}\left(W_R^{'}\vec{r}\right)\right) \tag{8}$$

Then, HIN2Vec adjusted the weights of $W_M^{'}\vec{m}, W_N^{'}\vec{n}$ and $W_R^{'}\vec{r}$ according to the gradients of $\log O_{m,n,r}(m,n,r)$ differentiated by $W_M^{'}\vec{m}, W_N^{'}\vec{n}$ and $W_R^{'}\vec{r}$, and thus maximized the objective function O. The specific definitions were as follows:

$$W_M^{t}\vec{m} := W_M^{t}\vec{m} + \frac{d\log O_{m,n,r}(m,n,r)}{dW_M^{t}\vec{m}} \tag{9}$$

$$W_N^{t}\vec{n} := W_N^{t}\vec{n} + \frac{d\log O_{m,n,r}(m,n,r)}{dW_N^{t}\vec{n}} \tag{10}$$

$$W_R^{t}\vec{r} := W_R^{t}\vec{r} + \frac{d\log O_{m,n,r}(m,n,r)}{dW_R^{t}\vec{r}} \tag{11}$$

Feature vector merging

With heterogeneous network representation learning method HeGAN and HIN2Vec based on the above generative adversarial network and meta-path random walk, we obtained two feature vector matrices, U and V, respectively. The final merging feature matrix X used in SMMA-HNRL was expressed with the following merging formula:

$$X = [U, V] \tag{12}$$

where [ ] represented the vector connect operation.

SMMA-HNRL model

In this study, we developed a novel model termed SMMA-HNRL to improve the performance of predicting potential SM-miRNA associations. The flowchart of SMMA-HNRL was shown in Figure 4.

Results

Evaluation metrics

The Recall, Precision, Accuracy, F1 Score, ROC curve with AUC (area under ROC curve) (Obuchowski and Bullen, 2018) value and PR curve with AUPR (area under PR curve) (Saito and Rehmsmeier, 2015) value were adopted as indicators for evaluating the performance of SMMA-HNRL. In contrast experiments, the average AUC values and the AUPR values of ten training sets of each model were calculated and the corresponding ROC curves and PR curves were drawn according to the results of 10-fold cross validation. Finally, the

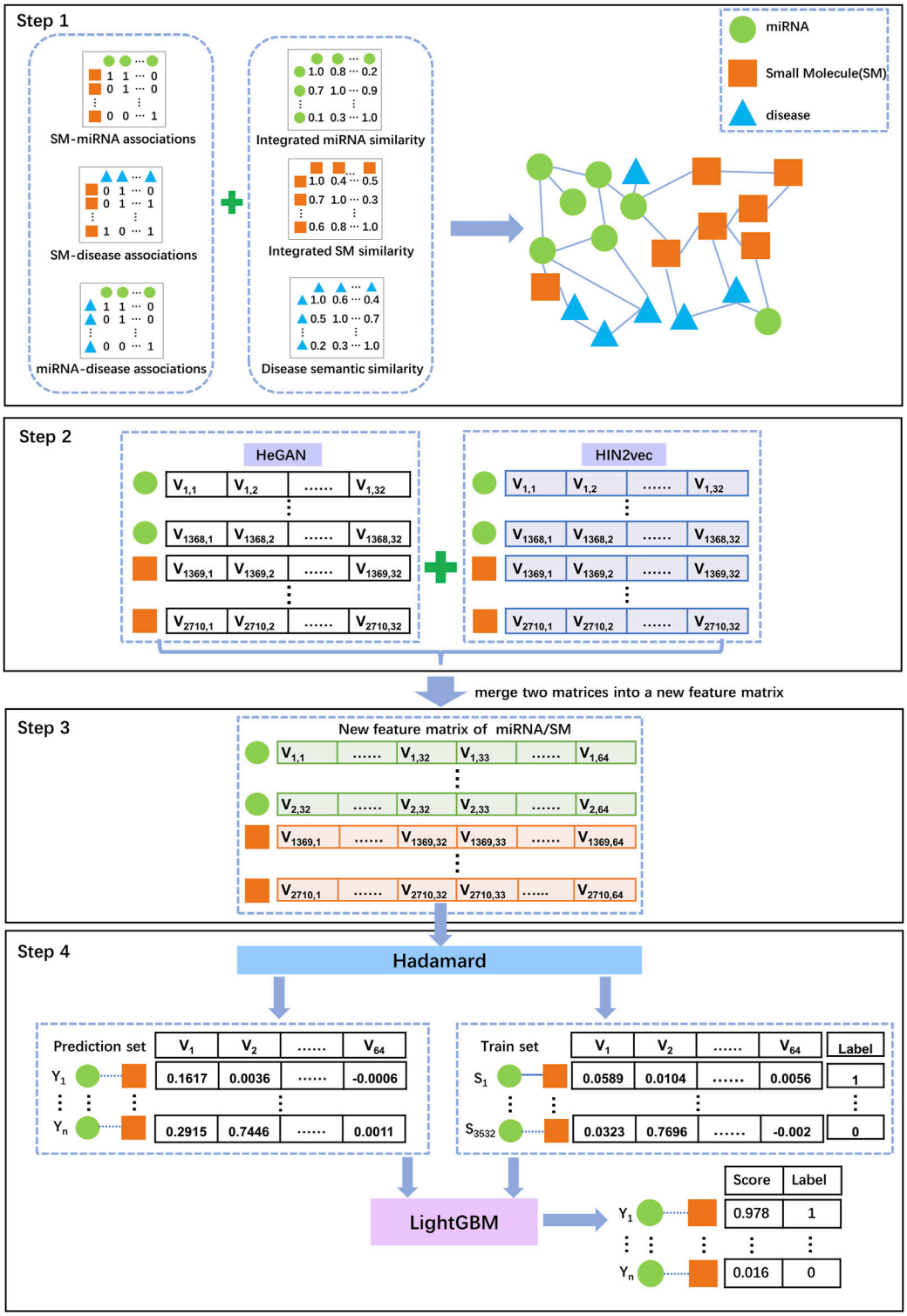

FIGURE 4
The flowchart of SMMA-HNRL model. Step 1: The associated data and similarity data obtained from different biological databases were preprocessed, a heterogeneous information network was constructed with three association networks (miRNA-SM, miRNA-disease, SM-disease) and three similarity networks (miRNA-miRNA, SM-SM, disease-disease) in our study. Step 2: With two different network representation learning algorithms, HeGAN and HIN2Vec, two feature matrices of the heterogeneous information networks were obtained. Step 3: Combining the feature vectors of miRNAs and SMs from two feature matrices, a merged feature matrix was finally obtained. Step 4: Hadamard function was adopted to convert SM and miRNA feature vectors into a feature vector for a SM-miRNA pair. The known SM-miRNA associations from the heterogeneous information network were chosen as training set for LightGBM classifier to predict potential SM-miRNA associations.

TABLE 1 The six evaluation metrics results of the three merging methods.

	Recall	Precision	Accuracy	F1 score	AUC	AUPR
Connection	0.9745	0.9563	0.9649	0.9652	0.9875	0.9885
Averaging	0.9513	0.9456	0.9482	0.9483	0.9828	0.9813
Multiplication	0.9439	0.9491	0.9465	0.9463	0.9814	0.9823

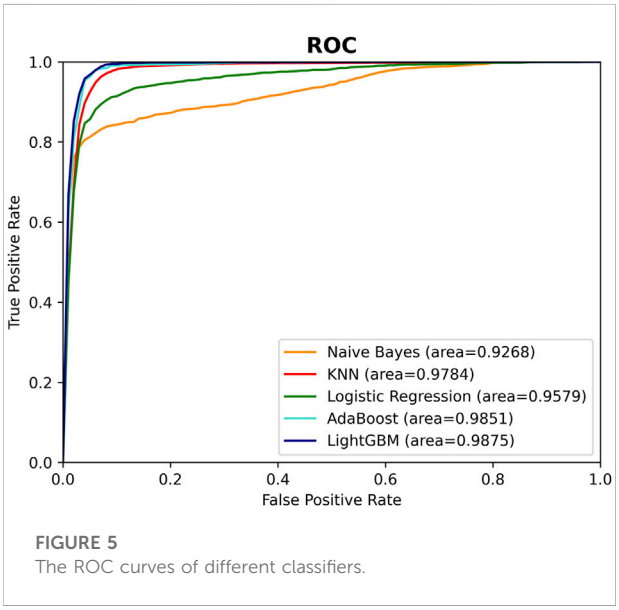

FIGURE 5
The ROC curves of different classifiers.

average values of all evaluation indicators were used to evaluate the model performance. For experiment details, please see Supplementary Material S2.

Comparison with other feature vector merging methods

After we obtained two sets of feature vectors of SM and miRNA by HeGAN and HIN2Vec, we conducted comparative experiments to evaluate the performances of different merging methods. Three merging methods (connection, averaging, and multiplication) were adopted to fuse two feature vectors of both SM and miRNA into one integrated vector. According to the experimental results, the connect operation had obtained the best performance. The detailed experimental results are shown in Table 1.

Classifier selection

After calculating of the SM-miRNA pair vectors, the problem of predicting potential SM-miRNA associations could be considered as a binary classification problem. In our study, 1766 pairs of SM-miRNA associations were downloaded from the SM2miR database as positive samples, and the same amount of SM-miRNA associations from all remaining combinations were randomly selected as negative samples. During the classifier selection, five different popular machine learning methods, Naive Bayes (NB) (Yang, 2018), Linear Regression (LR) (Maulud and Abdulazeez, 2020), K-Nearest Neighbor (KNN) (Zhang and Zhou, 2007), AdaBoost (Freund and Schapire, 1997) and LightGBM (Ke et al., 2017), were tested based on the merging feature vectors of the above samples, respectively. The performance of these five classifiers was evaluated with the Recall, Precision, Accuracy, F1 Score, AUC, and AUPR. Figure 5; Supplementary Material S3 illustrated the performance of these classifiers.

TABLE 2 Vector functions and AUC values of 10-fold cross validations.

Functions	Hadamard	Average	Minus	Absolute minus
Descriptions	$\overrightarrow{v_{1i}} * \overrightarrow{v_{2i}}$	$\frac{\overrightarrow{v_{1i}} + \overrightarrow{v_{2i}}}{2}$	$\overrightarrow{v_{1i}} - \overrightarrow{v_{2i}}$	$\lvert\overrightarrow{v_{1i}} - \overrightarrow{v_{2i}}\rvert$
AUC	0.9875	0.9829	0.9808	0.9701

Vector function selection for SM-miRNA pair

After gaining node vectors of SMs and miRNAs by the heterogeneous network representation learning algorithms, SM-miRNA pair vectors were subsequently calculated with vector functions. We study four commonly functions, Hadamard, Average, Minus and Absolute Minus (Deng et al., 2021), which merged one SM vector and one miRNA vector into one SM-miRNA pair vector. 10-fold cross validations of SMMA-HNRL with these four functions were employed in turn. Table 2 documented the descriptions of SM-miRNA pair vector functions and the corresponding AUC values of 10-fold cross validations. The experimental results demonstrated that Hadamard function outperformed the remaining three vector combinations. It could better remain the association between one SM vector and one miRNA vector. Therefore, Hadamard function was chosen as the SM-miRNA pair vector function for the following experiments.

TABLE 3 The six evaluation metrics results of different vector combinations.

	Recall	Precision	Accuracy	F1 score	AUC	AUPR
HeGAN16HIN2V16	0.9355	0.9550	0.9456	0.9450	0.9826	0.9839
HeGAN16HIN2V32	0.9575	0.9582	0.9578	0.9578	0.9865	0.9880
HeGAN16HIN2V64	0.9541	0.9575	0.9558	0.9557	0.9853	0.9882
HeGAN16HIN2V128	0.9541	0.9559	0.9550	0.9549	0.9858	0.9866
HeGAN32HIN2V16	0.9626	0.9514	0.9567	0.9569	0.9850	0.9861
HeGAN32HIN2V32	0.9745	0.9563	**0.9649**	**0.9652**	**0.9875**	0.9885
HeGAN32HIN2V64	0.9711	0.9529	0.9615	0.9619	0.9871	0.9886
HeGAN32HIN2V128	**0.9774**	0.9522	0.9640	0.9645	0.9868	0.9875
HeGAN64HIN2V16	0.9626	0.9514	0.9567	0.9569	0.9850	0.9861
HeGAN64HIN2V32	0.9632	0.9514	0.9570	0.9572	0.9859	0.9843
HeGAN64HIN2V64	0.9677	0.9586	0.9629	0.9631	0.9867	0.9873
HeGAN64HIN2V128	0.9694	0.9587	0.9638	0.9639	0.9870	0.9862
HeGAN128HIN2V16	0.9496	0.9546	0.9522	0.9520	0.9861	0.9868
HeGAN128HIN2V32	0.9615	**0.9606**	0.9609	0.9610	0.9874	**0.9890**
HeGAN128HIN2V64	0.9660	0.9580	0.9618	0.9619	0.9868	0.9879
HeGAN128HIN2V128	0.9643	0.9580	0.9609	0.9611	0.9869	0.9859

Parameter tuning

In SMMA-HNRL, HeGAN, and HIN2Vec were utilized for feature extraction respectively, which are both highly encapsulated representation learning models. In our study, most of the inner parameters of HeGAN and HIN2Vec were set to the recommended values, and the number of dimensions of features was treated as hyperparameters. The 16-dimensional, 32-dimensional, 64-dimensional and 128-dimensional topological feature vectors of SM nodes and miRNA nodes were calculated and merged respectively for gaining the optimal combination of feature vector dimensions. The results of 10-fold cross validation of different combinations were shown in Table 3. The numbers in the vector combination name represented feature vector dimensions, for example, HeGAN16HIN2V16 represented the integrated features combined with 16 dimensional features of HeGAN and 16 dimensional features of HIN2Vec. In Table 3, the best evaluation results of each metric were in bold. With a comprehensive consideration, HeGAN32HIN2V32 achieved the best results in the most evaluation metrics which was also marked in bold. Finally, HeGAN32HIN2V32 was selected as the final feature combination of SMMA-HNRL.

Ablation study

One of the significant characteristics of SMMA-HNRL is that the node feature vectors which were obtained from HeGAN and HIN2Vec are merged. To explore whether merging node features is effective for predicting SM-miRNA associations, we designed the ablation studies to evaluate the performance of the methods with only HeGAN feature vectors, only HIN2Vec feature vectors, and the merging of HeGAN feature vectors and HIN2Vec feature vectors. They were named as HeGAN32, HIN2V32, and HeGAN32HIN2V32 respectively. The results of different feature vector methods were shown in Table 4. It can be seen from Table 4 that HeGAN32HIN2V32 outperformed the other methods.

In order to confirm the hypothesis that adding disease association information can increase the information richness between miRNAs and SMs which would improve the accuracy of SM-miRNA association prediction, we designed two sets of experimental conditions for SMMA-HNRL. One set only contained heterogeneous information of three networks (SM-miRNA association network, integrated SM similarity network and integrated miRNA similarity network). There were only two kinds of nodes (miRNAs and SMs) in this heterogeneous network which was named as HIN-2N. The other set contained the heterogeneous information of all six networks which included three kinds of nodes (miRNAs, diseases and SMs) and was named as HIN-3N. The experimental results (AUC, AUPR) were shown in Figure 6. The AUC and AUPR values of HIN-3N are 0.9875 and 0.9885, which are both higher than those of HIN-2N. The comparison results for the remaining metrics are listed in Supplementary Material S4, and results showed they were all enhanced with different degree. This fully confirmed that introducing disease association information in our study is more reliable and more effective in predicting potential SM-miRNA associations.

TABLE 4 Ablation study results of different feature vector models.

	Recall	Precision	Accuracy	F1 score	AUC	AUPR
HeGAN32	0.9485	0.9471	0.9476	0.9476	0.9826	0.9822
HIN2V32	0.9417	0.9503	0.9462	0.9459	0.9825	0.9840
HeGAN32HIN2V32	0.9745	0.9563	0.9649	0.9652	0.9875	0.9885

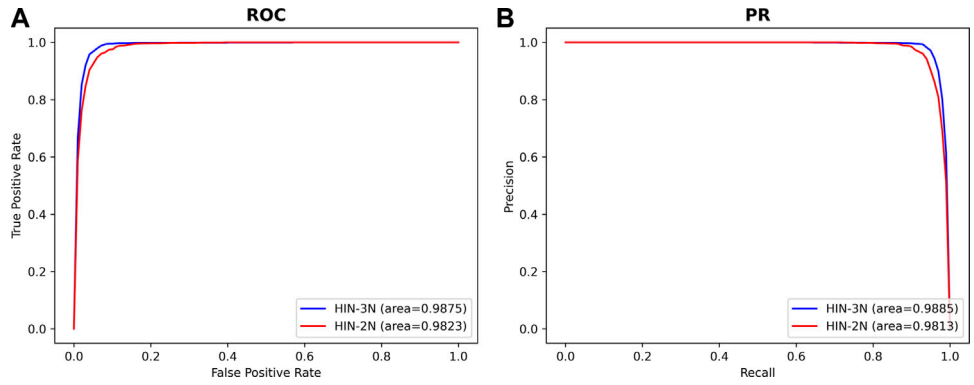

FIGURE 6
The ROC curves **(A)** and PR curves **(B)** on different heterogeneous networks.

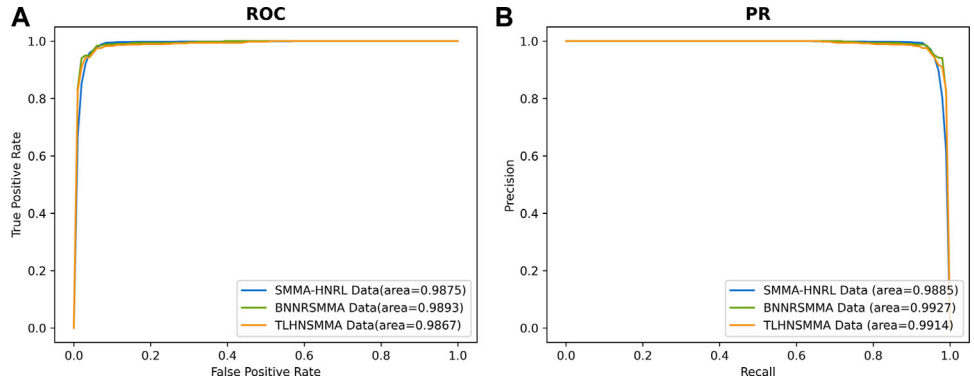

FIGURE 7
The ROC curves **(A)** and PR curves **(B)** of SMMA-HNRL in robustness testing.

Robustness testing

The robustness is the ability of one predictive model to maintain a stable performance on different scales and types of datasets. To evaluate the stability of SMMA-HNRL, two datasets were downloaded from BNNRSMMA (Chen et al., 2021) and TLHNSMMA (Qu et al., 2018) respectively as independent validation datasets. The BNNRSMMA dataset included 831 SMs, 541 miRNAs and contained 664 pairs of SM-miRNA associations. The TLHNSMMA dataset included 831 SMs, 541 miRNAs and 383 diseases. There were 664 SM-miRNA association pairs and 6233 miRNA-disease association pairs in it. The two datasets also include integrated similarities of SMs, miRNAs and diseases. In 10-

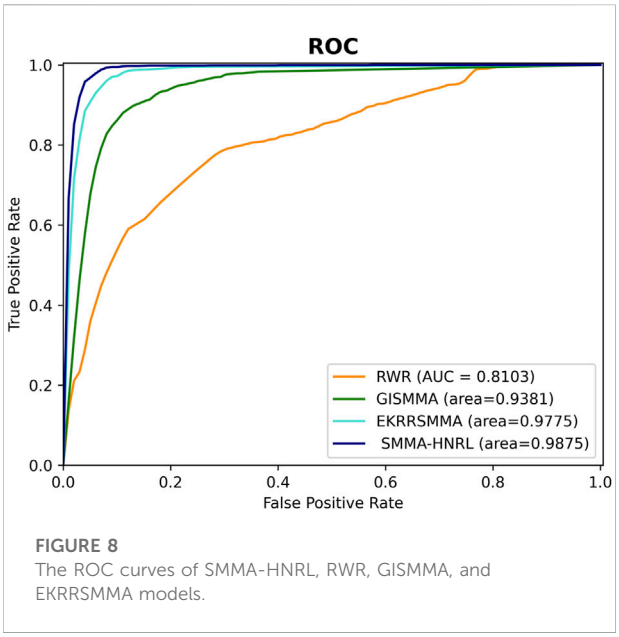

FIGURE 8
The ROC curves of SMMA-HNRL, RWR, GISMMA, and EKRRSMMA models.

fold cross validation of the two independent validation sets, all parameters of SMMA-HNRL were the same in both testing. AUC and AUPR values of SMMA-HNRL are shown in Figure 7, and the comparison of other metrics is shown in Supplementary Material S5. The results indicated that SMMA-HNRL achieved promising results on two sets and had power robust to different SM-miRNA datasets.

Independent test set validation

NoncoRNA database (Li et al., 2020) systematically recorded the information of ncRNAs with drug targets. Experimentally validated SM-miRNA interactions from NoncoRNA database were screened in our study and performed the same data preprocessing like those from SM2miR database. The data duplicated with SM2miR were eliminated with manual inspection. Finally, a total of 584 associations were obtained, including 272 miRNAs and 49 SMs. This data was not involved in the training of the model but as an independent test set to evaluate the generalization ability of SMMA-HNRL.

The ROC and PR curves of the experimental results of the independent test set which was exhibited in Supplementary Material S6. In the independent test set validation, the AUC and AUPR reached 0.9859 and 0.9859, respectively. This indicated that our model had a strong generalization capability, and the outperformance of SMMA-HNRL was not caused by overfittings.

Model contrast

To further demonstrate the predictive effectiveness of SMMA-HNRL, we compared SMMA-HNRL with the three state-of-the-art SM-miRNA association prediction models, EKRRSMMA (Wang et al., 2022), GISMMA (Guan et al., 2018) and RWR (Lv et al., 2015). In the contrast experiment, each model was trained and tested with the same datasets. The overview of datasets involved in each comparison model is exhibited in Supplementary Material S7. The prediction performance of the three models were performance with 10-fold cross validations, and the ROC curves were shown in Figure 8. It showed that SMMA-HNRL achieved AUC score of 0.9875, which outperformed RWR (AUC score: 0.8103), GISMMA (AUC score: 0.9381) and EKRRSMMA (AUC score: 0.9775). The merged feature vectors of SM nodes and miRNA nodes which were obtained with two different heterogeneous network representation learning algorithms (HeGAN and HIN2Vec) can improve prediction accuracy of potential SM-miRNA associations.

TABLE 5 Validation of the top 50 predicted miRNAs related to 5-FU (DB00544).

miRNA	Evidence	miRNA	Evidence
hsa-mir-135a	29,735,329	hsa-mir-139	27,173,050
hsa-mir-150	32,669,857	hsa-mir-133b	32,865,180
hsa-mir-181a	29,795,190	hsa-mir-134	34,168,463
hsa-mir-214	30,915,129	hsa-mir-130a	30,510,209
hsa-mir-29b	34,155,879	hsa-mir-130b	33,816,278
hsa-mir-30c	Unconfirmed	hsa-mir-30d	Unconfirmed
hsa-mir-320	25,446,103	hsa-mir-30b	Unconfirmed
hsa-mir-328	33,948,374	hsa-mir-29c	31,037,126
hsa-mir-425	32,158,234	hsa-mir-26a	29,719,405
hsa-mir-451	Unconfirmed	hsa-mir-28	30,762,286
hsa-mir-96	31,089,750	hsa-mir-24-1	Unconfirmed
hsa-mir-98	Unconfirmed	hsa-mir-212	32,862,489
hsa-mir-16-1	Unconfirmed	hsa-mir-22	25,449,431
hsa-mir-363	27,167,197	hsa-mir-221	27,726,102
hsa-mir-335	31,799,650	hsa-mir-223	Unconfirmed
hsa-mir-324	30,103,475	hsa-mir-20b	27,878,272
hsa-mir-326	26,239,225	hsa-mir-205	32,996,748
hsa-mir-424	33,793,771	hsa-mir-199b	32,580,513
hsa-mir-378	30,797,151	hsa-mir-100	Unconfirmed
hsa-mir-181a-2	Unconfirmed	hsa-let-7f-2	Unconfirmed
hsa-mir-181b	27,081,844	hsa-let-7	26,687,759
hsa-mir-186	Unconfirmed	hsa-let-7c	33,051,247
hsa-mir-193b	34,844,630	hsa-mir-1-1	Unconfirmed
hsa-mir-148b	Unconfirmed	hsa-mir-1-2	Unconfirmed
hsa-mir-144	32,162,886	hsa-mir-124-1	Unconfirmed

Case studies

To further evaluate the capability of SMMA-HNRL in practical applications, we conducted case studies with three

TABLE 6 Validation of the top 50 predicted miRNAs related to Cisplatin (DB00515).

miRNA	Evidence	miRNA	Evidence
hsa-mir-302b	26,623,722	hsa-mir-26a	26,458,859
hsa-mir-181a-2	34,815,714	hsa-mir-22	30,537,795
hsa-mir-186	32,284,740	hsa-mir-26b	31,686,855
hsa-mir-452	Unconfirmed	hsa-mir-30b	33,779,882
hsa-mir-9-3	Unconfirmed	hsa-mir-328	30,221,716
hsa-mir-191	32,803,782	hsa-mir-326	26,239,225
hsa-mir-29b-2	Unconfirmed	hsa-mir-320	Unconfirmed
hsa-mir-1-1	Unconfirmed	hsa-mir-144	31,017,720
hsa-mir-24	30,787,983	hsa-mir-145	31,821,542
hsa-mir-34b	33,720,323	hsa-mir-140	32,765,679
hsa-mir-193b	27,918,099	hsa-mir-134	Unconfirmed
hsa-mir-194	32,534,701	hsa-mir-132	31,906,769
hsa-mir-200a	32,256,108	hsa-mir-125a	33,777,215
hsa-mir-206	27,014,910	hsa-mir-127	Unconfirmed
hsa-mir-139	33,300,085	hsa-mir-210	30,957,179
hsa-mir-143	33,090,550	hsa-mir-212	Unconfirmed
hsa-mir-495	34,747,666	hsa-mir-193a	30,485,589
hsa-mir-7	33,072,745	hsa-mir-15a	26,314,859
hsa-mir-99b	30,984,249	hsa-mir-483	Unconfirmed
hsa-mir-10b	32,892,697	hsa-mir-486	32,527,702
hsa-mir-1	32,377,691	hsa-mir-99a	27,994,509
hsa-mir-122	27,874,954	hsa-let-7f-2	Unconfirmed
hsa-let-7a	29,565,706	hsa-let-7g	Unconfirmed
hsa-mir-92-1	Unconfirmed	hsa-let-7d	30,816,441
hsa-mir-25	27,743,413	hsa-let-7f	26,458,859

TABLE 7 Validation of the top 50 predicted miRNAs related to Imatinib (DB00619).

miRNA	Evidence	miRNA	Evidence
hsa-mir-34a	31,923,418	hsa-mir-26b	31,273,251
hsa-mir-155	30,459,357	hsa-let-7b	Unconfirmed
hsa-mir-21	28,190,319	hsa-let-7c	Unconfirmed
hsa-mir-221	30,516,071	hsa-mir-191	Unconfirmed
hsa-mir-145	Unconfirmed	hsa-mir-93	Unconfirmed
hsa-mir-125b-2	Unconfirmed	hsa-mir-99b	28,544,907
hsa-mir-204	Unconfirmed	hsa-mir-197	Unconfirmed
hsa-mir-107	Unconfirmed	hsa-mir-127	Unconfirmed
hsa-mir-92a-1	Unconfirmed	hsa-mir-27a	26,458,312
hsa-let-7i	28,512,058	hsa-mir-151a	28,544,907
hsa-mir-223	32,597,702	hsa-let-7e	33,066,614
hsa-mir-224	26,458,312	hsa-mir-424	25,697,481
hsa-mir-24	Unconfirmed	hsa-mir-30b	Unconfirmed
hsa-mir-18a	26,458,312	hsa-mir-222	30,396,237
hsa-mir-125b	Unconfirmed	hsa-mir-205	28,861,326
hsa-let-7a	Unconfirmed	hsa-mir-206	Unconfirmed
hsa-mir-148a	Unconfirmed	hsa-mir-373	Unconfirmed
hsa-mir-25	Unconfirmed	hsa-let-7	Unconfirmed
hsa-mir-27b	28,942,039	hsa-mir-494	28,533,480
hsa-mir-200a	28,942,039	hsa-mir-483	34,638,938
hsa-mir-19a	28,942,039	hsa-mir-34c	Unconfirmed
hsa-mir-1	Unconfirmed	hsa-mir-125b-1	Unconfirmed
hsa-mir-152	Unconfirmed	hsa-mir-200b	Unconfirmed
hsa-mir-15b	Unconfirmed	hsa-mir-22	Unconfirmed
hsa-mir-29b-1	31,923,418	hsa-mir-214	28,942,039

common SM drugs, 5-FU (DB00544), Cisplatin (DB00515) and Imatinib (DB00619), which were all closely related to human life and health.

First, the known SM-miRNA association vectors were used as positive samples and an equal amount of unknown associations that randomly generated was adopted as negative samples. Subsequently, SMMA-HNRL was trained with those samples. Then, miRNAs unrelated to the three SMs were screened from the heterogeneous network, and the SM-miRNA feature vectors were generated by Hadamard function. Finally, the feature vectors were input into the LightGBM classifier, and the probability scores were calculated. The predicted SM-miRNA associations were sorted in descending order by the probability scores. The potential associations were verified by manually reviewing the PubMed database for proving the effectiveness of SMMA-HNRL. The specific results were shown in the following table (Tables 5, 6, 7), and the full results for these three SMs are available in Supplementary Materials S8, S9, and S10.

5-FU, a pyrimidine analog, is a key chemotherapeutic drug in colorectal cancer (CRC) and has been implicated in the treatment of breast cancer. As an antimetabolite, it interferes with DNA synthesis by blocking the conversion of deoxyuridine to thymidylate by thymidylate synthase (Longley et al., 2003; Wigmore et al., 2010). Table 5 showed the top 50 miRNA associations associated with 5-FU, among the top 10, 8 miRNAs were confirmed by the literature, among the top 30, 23 miRNAs were confirmed by the literature, and among the top 50, 34 miRNAs were confirmed by the literature.

Cisplatin, the first metal-based anticancer drug, is widely used to treat various types of cancers, such as testicular cancer, ovarian cancer, lung cancer. It induces DNA damage by interacting with purine bases on DNA and eventually induces cancer cell apoptosis (Ghosh, 2019). Table 6 presented the top 50 miRNA associations associated with cisplatin, 6 miRNAs of the top 10, 25 miRNAs of the top 30 and 37 miRNAs of the top 50 were documented confirmed by the literature.

Imatinib is a potent drug for chronic myeloid leukemia, which inhibits the rapid division of cancer cells by inhibiting specific tyrosine kinases (Peng et al., 2005). Table 7 showed the

top 50 miRNA associations associated with imatinib, 5 miRNAs of the top 10, 15 miRNAs of the top 30 and 22 miRNAs of the top 50 were confirmed by the literature.

In summary, through the case studies of 5-FU, Cisplatin and Imatinib, the majority of novel associations with the highest probability has been confirmed by the PubMed literatures, and it is enough to illustrate the outstanding performance of SMMA-HNRL in predicting potential SM-miRNA associations.

Discussion

Numerous studies proved that many human complex diseases are closely related to the dysregulations of related key miRNAs, and miRNAs have been recognized as a potential class of drug targets. Predicting novel SM-miRNA associations is important to help researchers find effective drugs, understand the molecular basis of diseases and reduce experimental costs. Nowadays, it has become a trend to construct heterogeneous networks to predict potential SM-miRNA associations by integrating multiple biological entities. In this work, we proposed a novel model SMMA-HNRL based on an integrated heterogeneous network representation learning algorithm. By building a heterogeneous information network with the HeGAN algorithm based on the generative adversarial network and the HIN2Vec algorithm based on the random walk of the meta-path, richer feature information of the heterogeneous network was accessed, which overcame the data sparsity due to few known associations. Validated by the experiments, SMMA-HNRL exhibited high robustness. Compared with three state-of-the-art predicting models, our model achieved the best performance under the same dataset evaluated by 10-fold cross validation. Case studies of three common drugs showed that the model had good application significance. Otherwise, SMMA-HNRL can be regarded as an open framework, and users can adopt more heterogeneous information related with SM-miRNA association prediction for improving prediction accuracy.

Although SMMA-HNRL has achieved satisfactory results in predicting potential SM-miRNA associations, there is still room for improvement in the experiments. First, due to the current biological limitations, we do not exactly find negative samples of SM-miRNA associations, so future research will introduce more effective negative sample screening methods. Second, although this study introduced multi-source heterogeneous data for network construction, it still cannot fully reflect the comprehensive and complex interaction network. The more data integrated, the higher the accuracy and robustness of the model will be. In the future, more biological data will be introduced for processing, such as lncRNA and gene related data.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary Material, further inquiries can be directed to the corresponding author.

Author contributions

JL conceived, designed the study, HL and YW developed the algorithm and performed the statistical analysis, HL, ZL, and BW wrote the codes. HL drafted the original manuscript; JL revised the manuscript. All authors read and approved the final manuscript.

Funding

This work is supported by the National Natural Science Foundation of China under grants No. 62072154 and 62202330.

Acknowledgments

We thank members in our groups for their valuable discussions.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2022.1079053/full#supplementary-material

References

Azad, A. K. M., Dinarvand, M., Nematollahi, A., Swift, J., Lutze-Mann, L., and Vafaee, F. (2021). A comprehensive integrated drug similarity resource for *in-silico* drug repositioning and beyond. *Brief. Bioinform.* 22 (3), bbaa126. doi:10.1093/bib/bbaa126

Bartel, D. P. (2004). MicroRNAs: Genomics, biogenesis, mechanism, and function. *Cell* 116 (2), 281–297. doi:10.1016/s0092-8674(04)00045-5

Chen, X., Guan, N. N., Sun, Y. Z., Li, J. Q., and Qu, J. (2018). MicroRNA-small molecule association identification: From experimental results to computational models. *Brief. Bioinform.* doi:10.1093/bib/bby098

Chen, X., Sun, Y. Z., Zhang, D. H., Li, J. Q., Yan, G. Y., An, J. Y., et al. (2017b2017). *Nrdtd: A database for clinically or experimentally supported non-coding RNAs and drug targets associations*. Oxford: Database. doi:10.1093/database/bax057

Chen, X., Xie, D., Zhao, Q., and You, Z. H. (2019). MicroRNAs and complex diseases: From experimental results to computational models. *Brief. Bioinform.* 20 (2), 515–539. doi:10.1093/bib/bbx130

Chen, X., Xie, W. B., Xiao, P. P., Zhao, X. M., and Yan, H. (2017a). mTD: A database of microRNAs affecting therapeutic effects of drugs. *J. Genet. Genomics* 44 (5), 269–271. doi:10.1016/j.jgg.2017.04.003

Chen, X., Zhou, C., Wang, C. C., and Zhao, Y. (2021). Predicting potential small molecule-miRNA associations based on bounded nuclear norm regularization. *Brief. Bioinform.* 22 (6), bbab328. doi:10.1093/bib/bbab328

Cheng, A. M., Byrom, M. W., Shelton, J., and Ford, L. P. (2005). Antisense inhibition of human miRNAs and indications for an involvement of miRNA in cell growth and apoptosis. *Nucleic Acids Res.* 33 (4), 1290–1297. doi:10.1093/nar/gki200

Cheng, C. J., Bahal, R., Babar, I. A., Pincus, Z., Barrera, F., Liu, C., et al. (2015). MicroRNA silencing for cancer therapy targeted to the tumour microenvironment. *Nature* 518 (7537), 107–110. doi:10.1038/nature13905

Croce, C. M., and Calin, G. A. (2005). miRNAs, cancer, and stem cell division. *Cell* 122 (1), 6–7. doi:10.1016/j.cell.2005.06.036

Davis, A. P., Grondin, C. J., Johnson, R. J., Sciaky, D., King, B. L., McMorran, R., et al. (2017). The comparative toxicogenomics database: Update 2017. *Nucleic Acids Res.* 45 (D1), D972–D978. doi:10.1093/nar/gkw838

Deng, L., Li, W., and Zhang, J. (2021). LDAH2V: Exploring meta-paths across multiple networks for lncRNA-disease association prediction. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 18 (4), 1572–1581. doi:10.1109/TCBB.2019.2946257

Dong, Y., Chawla, N. V., and Swami, A. (2017). "metapath2vec: Scalable representation learning for heterogeneous networks," in Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (Halifax, NSCanada: Association for Computing Machinery), 135–144. doi:10.1145/3097983.3098036

Freund, Y., and Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. Syst. Sci.* 55 (1), 119–139. doi:10.1006/jcss.1997.1504

Fu, T.-y., Lee, W.-C., and Lei, Z. (2017). "HIN2Vec: Explore meta-paths in heterogeneous information networks for representation learning," in Proceedings of the 2017 ACM on Conference on Information and Knowledge Management (Singapore, Singapore: Association for Computing Machinery). doi:10.1145/3132847.3132953

Ghosh, S. (2019). Cisplatin: The first metal based anticancer drug. *Bioorg. Chem.* 88, 102925. doi:10.1016/j.bioorg.2019.102925

Grover, A., and Leskovec, J. (2016). "node2vec: Scalable feature learning for networks," in Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining (San Francisco, California, USA: Association for Computing Machinery), 855–864. doi:10.1145/2939672.2939754

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. doi:10.3389/fphar.2018.01152

Hu, B., Fang, Y., and Shi, C. (2019). "Adversarial learning on heterogeneous information networks," in Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (Anchorage, AK, USA: Association for Computing Machinery), 120–129. doi:10.1145/3292500.3330970

Huang, Z., Shi, J., Gao, Y., Cui, C., Zhang, S., Li, J., et al. (2019). HMDD v3.0: A database for experimentally supported human microRNA-disease associations. *Nucleic Acids Res.* 47 (D1), D1013–D1017. doi:10.1093/nar/gky1010

Ji, B. Y., You, Z. H., Jiang, H. J., Guo, Z. H., and Zheng, K. (2020). Prediction of drug-target interactions from multi-molecular network based on LINE network representation method. *J. Transl. Med.* 18 (1), 347. doi:10.1186/s12967-020-02490-x

Jiang, W., Chen, X., Liao, M., Li, W., Lian, B., Wang, L., et al. (2012). Identification of links between small molecules and miRNAs in human cancers based on transcriptional responses. *Sci. Rep.* 2, 282. doi:10.1038/srep00282

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., et al. (2017). "LightGBM: A highly efficient gradient boosting decision tree," in Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach (California, USA: Curran Associates Inc), 3149–3157. doi:10.5555/3294996.3295074

Lee, R. C., Feinbaum, R. L., and Ambros, V. (1993). The *C. elegans* heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell* 75 (5), 843–854. doi:10.1016/0092-8674(93)90529-y

Li, J., Lei, K., Wu, Z., Li, W., Liu, G., Liu, J., et al. (2016). Network-based identification of microRNAs as potential pharmacogenomic biomarkers for anticancer drugs. *Oncotarget* 7 (29), 45584–45596. doi:10.18632/oncotarget.10052

Li, J., Li, J., Kong, M., Wang, D., Fu, K., and Shi, J. (2021). Svdnvlda: Predicting lncRNA-disease associations by singular value decomposition and node2vec. *BMC Bioinforma.* 22 (1), 538. doi:10.1186/s12859-021-04457-1

Li, J., Zhang, S., Wan, Y., Zhao, Y., Shi, J., Zhou, Y., et al. (2019). MISIM v2.0: A web server for inferring microRNA functional similarity based on microRNA-disease associations. *Nucleic Acids Res.* 47 (W1), W536–W541. doi:10.1093/nar/gkz328

Li, L., Wu, P., Wang, Z., Meng, X., Zha, C., Li, Z., et al. (2020). NoncoRNA: A database of experimentally supported non-coding RNAs and drug targets in cancer. *J. Hematol. Oncol.* 13 (1), 15. doi:10.1186/s13045-020-00849-7

Liu, X., Wang, S., Meng, F., Wang, J., Zhang, Y., Dai, E., et al. (2013). SM2miR: A database of the experimentally validated small molecules' effects on microRNA expression. *Bioinformatics* 29 (3), 409–411. doi:10.1093/bioinformatics/bts698

Liu, Z., Sall, A., and Yang, D. (2008). MicroRNA: An emerging therapeutic target and intervention tool. *Int. J. Mol. Sci.* 9 (6), 978–999. doi:10.3390/ijms9060978

Longley, D. B., Harkin, D. P., and Johnston, P. G. (2003). 5-fluorouracil: Mechanisms of action and clinical strategies. *Nat. Rev. Cancer* 3 (5), 330–338. doi:10.1038/nrc1074

Lv, Y., Wang, S., Meng, F., Yang, L., Wang, Z., Wang, J., et al. (2015). Identifying novel associations between small molecules and miRNAs based on integrated molecular networks. *Bioinformatics* 31 (22), 3638–3644. doi:10.1093/bioinformatics/btv417

Maulud, D., and Abdulazeez, A. M. (2020). A review on linear regression comprehensive in machine learning. *J. Appl. Sci. Technol. Trends* 1 (4), 140–147. doi:10.38094/jastt1457

Miska, E. A. (2005). How microRNAs control cell division, differentiation and death. *Curr. Opin. Genet. Dev.* 15 (5), 563–568. doi:10.1016/j.gde.2005.08.005

Obuchowski, N. A., and Bullen, J. A. (2018). Receiver operating characteristic (ROC) curves: Review of methods with applications in diagnostic medicine. *Phys. Med. Biol.* 63 (7), 07tr01. doi:10.1088/1361-6560/aab4b1

Peng, B., Lloyd, P., and Schran, H. (2005). Clinical pharmacokinetics of imatinib. *Clin. Pharmacokinet.* 44 (9), 879–894. doi:10.2165/00003088-200544090-00001

Qu, J., Chen, X., Sun, Y. Z., Li, J. Q., and Ming, Z. (2018). Inferring potential small molecule-miRNA association based on triple layer heterogeneous network. *J. Cheminform.* 10 (1), 30. doi:10.1186/s13321-018-0284-9

Rossi, J. J. (2009). New hope for a microRNA therapy for liver cancer. *Cell* 137 (6), 990–992. doi:10.1016/j.cell.2009.05.038

Saito, T., and Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS One* 10 (3), e0118432. doi:10.1371/journal.pone.0118432

Sayed, D., and Abdellatif, M. (2011). MicroRNAs in development and disease. *Physiol. Rev.* 91 (3), 827–887. doi:10.1152/physrev.00006.2010

Shen, C., Luo, J., Ouyang, W., Ding, P., and Wu, H. (2020). Identification of small molecule-miRNA associations with graph regularization techniques in heterogeneous networks. *J. Chem. Inf. Model.* 60 (12), 6709–6721. doi:10.1021/acs.jcim.0c00975

Shivdasani, R. A. (2006). MicroRNAs: Regulators of gene expression and cell differentiation. *Blood* 108 (12), 3646–3653. doi:10.1182/blood-2006-01-030015

Stern-Ginossar, N., Elefant, N., Zimmermann, A., Wolf, D. G., Saleh, N., Biton, M., et al. (2007). Host immune system gene targeting by a viral miRNA. *Science* 317 (5836), 376–381. doi:10.1126/science.1140956

Tang, J., Qu, M., Wang, M., Zhang, M., Yan, J., and Mei, Q. (2015). "Line: Large-Scale information network embedding," in Proceedings of the 24th International

Conference on World Wide Web (Florence, Italy: International World Wide Web Conferences Steering Committee), 1067–1077. doi:10.1145/2736277.2741093

Thafar, M. A., Olayan, R. S., Albaradei, S., Bajic, V. B., Gojobori, T., Essack, M., et al. (2021). DTi2Vec: Drug-target interaction prediction using network embedding and ensemble learning. *J. Cheminform.* 13 (1), 71. doi:10.1186/s13321-021-00552-w

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 (4), 1668–1679. doi:10.1021/acs.jcim.9b00129

Wang, C. C., Zhu, C. C., and Chen, X. (2022). Ensemble of kernel ridge regression-based small molecule-miRNA association prediction in human disease. *Brief. Bioinform.* 23 (1), bbab431. doi:10.1093/bib/bbab431

Wang, D., Wang, J., Lu, M., Song, F., and Cui, Q. (2010). Inferring the human microRNA functional similarity and functional network based on microRNA-associated diseases. *Bioinformatics* 26 (13), 1644–1650. doi:10.1093/bioinformatics/btq241

Wigmore, P. M., Mustafa, S., El-Beltagy, M., Lyons, L., Umka, J., and Bennett, G. (2010). Effects of 5-FU. *Adv. Exp. Med. Biol.* 678, 157–164. doi:10.1007/978-1-4419-6306-2_20

Wishart, D. S., Feunang, Y. D., Guo, A. C., Lo, E. J., Marcu, A., Grant, J. R., et al. (2018). DrugBank 5.0: A major update to the DrugBank database for 2018. *Nucleic Acids Res.* 46 (D1), D1074–D1082. doi:10.1093/nar/gkx1037

Xie, W. B., Yan, H., and Zhao, X. M. (2019). EmDL: Extracting miRNA-drug interactions from literature. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 16 (5), 1722–1728. doi:10.1109/TCBB.2017.2723394

Xu, P., Guo, M., and Hay, B. A. (2004). MicroRNAs and the regulation of cell death. *Trends Genet.* 20 (12), 617–624. doi:10.1016/j.tig.2004.09.010

Xuan, P., Han, K., Guo, M., Guo, Y., Li, J., Ding, J., et al. (2013). Prediction of microRNAs associated with human diseases based on weighted k most similar neighbors. *PLoS One* 8 (8), e70204. doi:10.1371/journal.pone.0070204

Yang, F. J. (2018). "An implementation of naive Bayes classifier," in 2018 International Conference on Computational Science and Computational Intelligence (CSCI), 301–306. doi:10.1109/CSCI46756.2018.00065

Yang, Y., Fu, X., Qu, W., Xiao, Y., and Shen, H. B. (2018). MiRGOFS: A GO-based functional similarity measurement for miRNAs, with applications to the prediction of miRNA subcellular localization and miRNA-disease association. *Bioinformatics* 34 (20), 3547–3556. doi:10.1093/bioinformatics/bty343

Yue, X., Wang, Z., Huang, J., Parthasarathy, S., Moosavinasab, S., Huang, Y., et al. (2020). Graph embedding on biomedical networks: Methods, applications and evaluations. *Bioinformatics* 36 (4), 1241–1251. doi:10.1093/bioinformatics/btz718

Zhang, M.-L., and Zhou, Z.-H. (2007). ML-KNN: A lazy learning approach to multi-label learning. *Pattern Recognit. DAGM.* 40 (7), 2038–2048. doi:10.1016/j.patcog.2006.12.019

Zhang, W., Yue, X., Lin, W., Wu, W., Liu, R., Huang, F., et al. (2018). Predicting drug-disease associations by using similarity constrained matrix factorization. *BMC Bioinforma.* 19 (1), 233. doi:10.1186/s12859-018-2220-4

Zhao, Q., Yu, H., Ji, M., Zhao, Y., and Chen, X. (2019). Computational model development of drug-target interaction prediction: A review. *Curr. Protein Pept. Sci.* 20 (6), 492–494. doi:10.2174/1389203720666190123164310

Zhu, S., Bing, J., Min, X., Lin, C., and Zeng, X. (2018). Prediction of drug-gene interaction by using Metapath2vec. *Front. Genet.* 9, 248. doi:10.3389/fgene.2018.00248