JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

pubs.acs.org/jcim Article

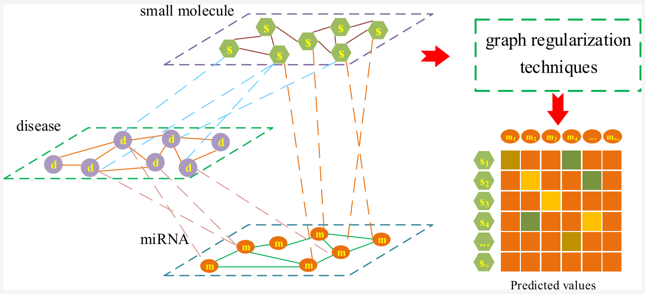
Identification of Small Molecule–miRNA Associations with Graph Regularization Techniques in Heterogeneous Networks

Cong Shen, Jiawei Luo,* Wenjue Ouyang, Pingjian Ding, and Hao Wu

Cite This: *J. Chem. Inf. Model.* 2020, 60, 6709–6721 Read Online

ACCESS | Metrics & More | Article Recommendations | SI Supporting Information

ABSTRACT: MicroRNAs (miRNAs) are significant regulators of post-transcriptional levels and have been confirmed to be targeted by small molecule (SM) drugs. It is a novel insight to treat human diseases and accelerate drug discovery by targeting miRNA with small molecules. Computational approaches for discovering novel small molecule–miRNA associations by integrating more heterogeneous network information provide a new idea for the multiple node association prediction between small molecule–miRNA and small molecule–disease associations at a system level. In this study, we proposed a new computational model based on graph regularization techniques in heterogeneous networks, called identification of small molecule–miRNA associations with graph regularization techniques (SMMARTs), to discover potential small molecule–miRNA associations. The novelty of the model lies in the fact that the association score of a small molecule–miRNA pair is calculated by an iterative method in heterogeneous networks that incorporates small molecule–disease associations and miRNA–disease associations. The experimental results indicate that SMMART has better performance than several state-of-the-art methods in inferring small molecule–miRNA associations. Case studies further illustrate the effectiveness of SMMART for small molecule–miRNA association prediction.



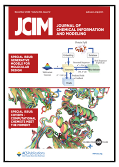
1. INTRODUCTION

MicroRNAs (miRNAs) are noncoding RNAs (ncRNAs) that play important roles in many biological processes, such as cell differentiation, proliferation, and apoptosis.[1,2] Structurally, each miRNA contains 19–24 nucleotides. Functionally, miRNAs can regulate gene expression through a sequence-specific approach.[3−5] Because miRNAs are ubiquitous in pathological processes, they have been suggested to become potential drug targets,[6−8] and the number of research hotspots in miRNA-centered computational biology has increased.[9−11] Recent studies have found that mature miRNAs and their precursors can be targeted by drugs.[6,12−14]

A recent study estimated that developing a new Food and Drug Administration (FDA)-approved drug cost an average of 2.6 billion in 2015, up from just 802 million in 2003.[15] Thus, it is a time-consuming and expensive process to develop a new drug. Modern drug discovery aims to speed up the research steps and thus reduces cost by leveraging computational tools on drug discovery. In short, molecular compounds are filtered through a progressive series of tests, which determine their properties, effectiveness, and toxicity for later stages. The computational method is increasingly being used to better predict molecular properties in early stages, which can significantly reduce the load of later processes (e.g., clinical trials) and save tons of resources as well as time.[16−18] The prediction of the association between small molecules (SMs) and miRNA is one of the important methods of drug discovery. Predicting the association between small molecules and miRNA based on computational models can reduce the workload of wet laboratory experiments and save resources. Meanwhile, predicting the association between drugs and diseases also is an important research area of drug discovery.[19−21] If the association of drug–miRNA (a target) and drug–disease can be completed in the same system, there may be a significant improvement in predicted performance. Hence, it is necessary to propose some novel calculative models to systematically analyze the associations between some drugs and miRNAs to accelerate the research of pharmacogenomics.

As mentioned above, there have been some computational methods to predict the associations between small molecules and miRNAs in previous studies.[22] For instance, feature-based models are the primary types. Velagapudi et al.[23] described a method called Inforna that designs small molecules only from the sequence level of miRNA. Based on the above studies, Velagapudi et al.[23] employed a novel version termed Inforna

Received: August 20, 2020
Published: November 9, 2020



https://dx.doi.org/10.1021/acs.jcim.0c00975

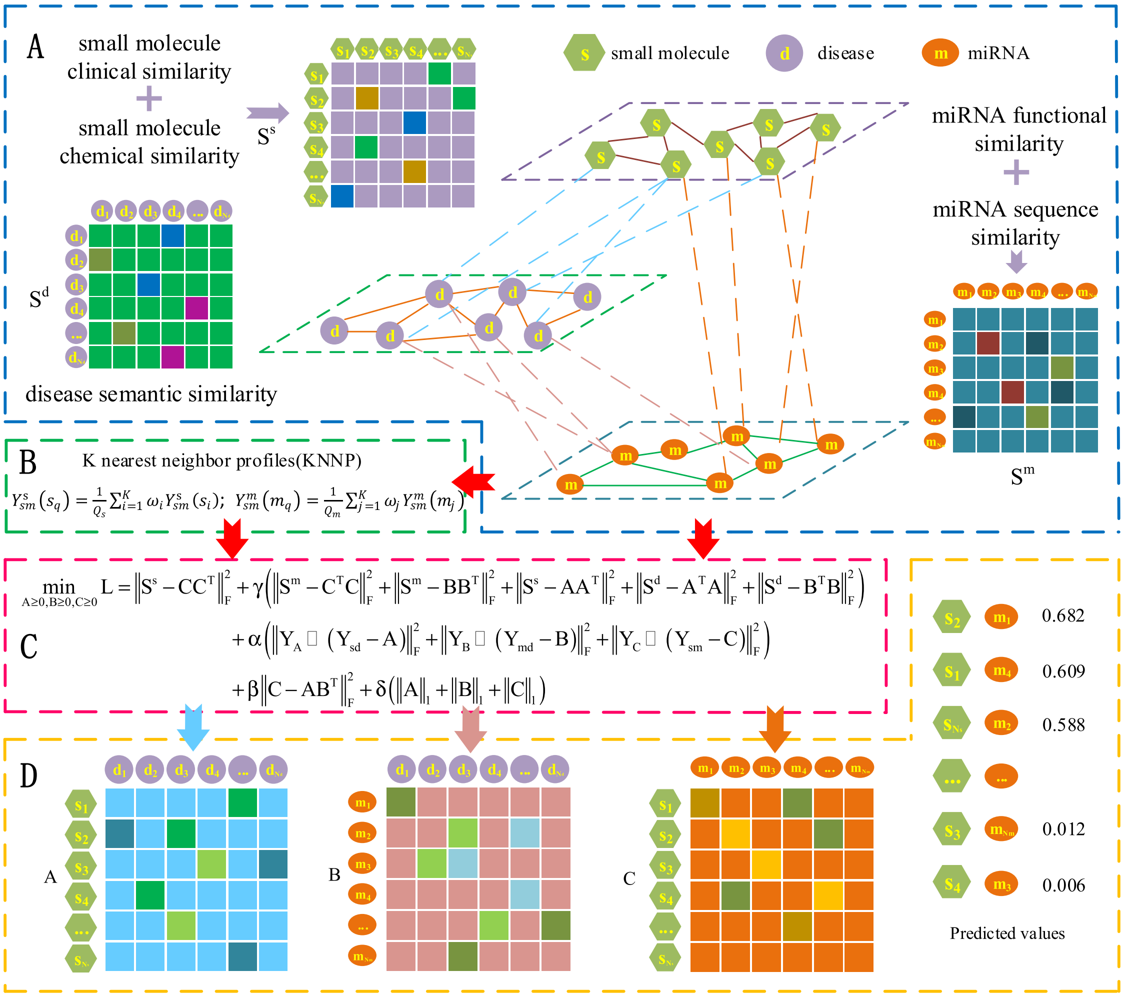

A
small molecule clinical similarity
+
small molecule chemical similarity
S^s
s_1 s_2 s_3 s_4 ... s_{N_s}
small molecule
disease
miRNA
miRNA functional similarity
+
miRNA sequence similarity
S^d
d_1 d_2 d_3 d_4 ... d_{N_d}
disease semantic similarity
S^m
m_1 m_2 m_3 m_4 ... m_{N_m}
B
K nearest neighbor profiles(KNNP)
$Y_{sm}^{s}(s_q)=\frac{1}{Q_s}\sum_{i=1}^{K}\omega_i Y_{sm}^{s}(s_i);\ Y_{sm}^{m}(m_q)=\frac{1}{Q_m}\sum_{j=1}^{K}\omega_j Y_{sm}^{m}(m_j)$
C
$\min_{A\geq0,B\geq0,C\geq0} L=\|S^s-CC^T\|_F^2+\gamma\left(\|S^m-C^TC\|_F^2+\|S^m-BB^T\|_F^2+\|S^s-AA^T\|_F^2+\|S^d-A^TA\|_F^2+\|S^d-B^TB\|_F^2\right)$
$+\alpha\left(\|Y_A\square(Y_{sd}-A)\|_F^2+\|Y_B\square(Y_{md}-B)\|_F^2+\|Y_C\square(Y_{sm}-C)\|_F^2\right)$
$+\beta\|C-AB^T\|_F^2+\delta\left(\|A\|_1+\|B\|_1+\|C\|_1\right)$
D
A
B
C
s_2 m_1 0.682
s_1 m_4 0.609
s_{N_s} m_2 0.588
... ...
s_3 m_{N_m} 0.012
s_4 m_3 0.006
Predicted values

Figure 1. Pipeline of SMMART for discovering unknown associations between small molecules and miRNAs. (A) Biological heterogeneous network constructed by the associations of small molecules, miRNAs, and disease nodes and their similarities. (B) Application of the KNNP model to improve the prediction performance of the model. (C) Application of graph regularization technology to fuse multisource information and predict the association of potential small molecules and miRNAs. (D) Prediction matrix of small molecule−miRNA association, the prediction matrix of small molecule−disease association, and the prediction matrix of miRNA−disease association. Finally, we use the correlation matrix score of small molecule−miRNA for the SMMART model.

2.0,[24] which is an online web server. In addition to the way in which sequences are transformed into features, there are also some ways to extract features from the literature. Xie et al.[25] developed a new text mining method, called EmDL, to discover the associations between small molecule drugs and miRNAs of affecting drug efficacy from the literature. Later, the authors compiled the association data of small molecules and miRNAs found in the literature into a database.[26] Similarly, a noncoding RNA and drug association database NRDTD[27] is proposed. The NRDTD data set mainly contains 97 noncoding RNAs and 96 drugs. Another important method to infer small molecule-associated miRNAs is developed by extracting features from the similarity data of miRNAs and small molecules. Wang et al.[28] proposed a novel calculation framework (RFSMMA) using random forest model to predict potential small molecule−miRNA associations based on extracting features from the similarity matrix of small molecules and miRNAs. The feature-based method has achieved good results in small molecule−miRNA association prediction, but most feature-based methods are supervised methods. However, sampling is a key step in the supervised method; the sampling method is one of the important methods to achieve outstanding performance in the imbalanced data set. In this study, it is almost impossible to obtain negative samples of small molecule-related miRNAs. Therefore, many machine learning methods, such as supervised learning, etc., selected negative samples from neutral samples for training models, which may add noise into the training set and affect the prediction performance. Furthermore, it is hard to decide which sampling method is suitable for machine learning methods of predicting small molecule−miRNA associations. Overall, predicting small molecule-associated miRNAs without selection of negative samples is a new and promising perspective.

In contrast, network-based models are a type of predictive model that do not require sampling for prioritizing the small molecule−miRNA associations. Lv et al.[29] utilized random walk with restart (RWR) to predict potential miRNA targets of 831 small molecules in a heterogeneous network. Considering the integration of neighborhood information, Li et al.[30] developed a network-based framework, termed SMiR-NBI, to use miRNAs as potential biomarkers for characterization of anticancer drug response in a heterogeneous network, which included drugs, genes, and miRNAs. Guan et al.[31] introduced a prediction method, called GISMMA, which used the graphlet interaction to predict unknown small molecule−miRNA associations. Zhao et al.[32] presented a model (SNMFSMMA) based on symmetric non-negative matrix factorization. Qu et al.[33] proposed a HeteSim-based inference model for predicting novel small molecule−miRNA associations, called HSSMMA, which implements a path-based measurement method of HeteSim on a heterogeneous network. Due to the sparsity of network data, Yin et al.[34] proposed a computational framework to predict small molecule−miRNA associations based on sparse learning and heterogeneous graph inference. However, most of the existing methods for discovering associations between small molecules and miRNAs are limited to only miRNA similarity networks, small molecule similarity networks, or bipartite small molecule−miRNA models. Meanwhile, miRNAs are promising therapeutic targets for complex diseases because small molecule drugs can regulate the expression of disease-related miRNAs.[35] Then, if both drug−miRNA (a target) and drug−disease relationships are considered for drug repositioning, this may be a new perspective for computer-aided drug design. Qu et al.[36] developed a computational model (TLHNSMMA) to uncover potential small molecule−miRNA associations based on a triple layer heterogeneous network containing small molecules, miRNAs, and diseases. Wang et al.[37] used cross-layer dependency inference on multilayered networks to predict small molecule-associated miRNAs (CLDISMMA), which constructed multilayered networks composed of SMs, miRNAs, and diseases. These methods have achieved good performance in the prediction of small molecule−miRNA relationships and may provide great help for the downstream wet experiment of small molecule-related miRNA relationship prediction. To fully consider the nodes and their associations, and effectively fuse multisource information in heterogeneous networks, we utilize graph regularization technology to predict the associations between small molecules and miRNAs by correlating small molecule−miRNA associations with small molecule−disease at a system level.

In this study, we described a novel framework, termed identification of small molecule−miRNA associations with graph regularization techniques (SMMARTs), to systematically infer unknown associations between small molecules and miRNAs. The SMMART model fully exploits the topological information of heterogeneous networks that consider small molecule−miRNA associations and small molecule−disease associations at a system level to achieve a better predictive performance of small molecule−miRNA association. To illustrate the performance of the SMMART model, we use the 5-fold cross-validation method to compare with other models. At the same time, to identify robustness, we use 50−90% of the known association data as the training set and use the remaining data as the test set to verify performance. Finally, case studies are used to further analyze the ability of the SMMART model to discover new small molecule−miRNA associations.

2. METHODS AND MATERIALS

2.1. Method Overview. We develop a novel method called SMMART to discover small molecule−miRNA associations. Figure 1 shows the overall workflow of our framework, which consists of three main steps. First, we construct a heterogeneous network, which includes three types of nodes (small molecules, miRNAs, and diseases). Second, the graph regularization technique is used to construct the predictive model (SMMART) by combining similarity prior knowledge, internode associations, and topological characteristics. Third, the association score between small molecule and miRNA is obtained based on the results of our framework.

2.2. Computation and Representation of Multisource Data. *2.2.1. Representation of Small Molecule Similarities.* *2.2.1.1. Clinical Similarity.* As the chemical structure, pharmacological effects, and therapeutic effects of small molecules are incorporated in the ATC code,[38] the small molecule clinical similarity calculated by the ATC code in many studies has been widely used to improve the performance of drug target prediction[39] and drug combination study.[40,41] Thus, we download the ATC codes from DrugBank[42] and use ATC codes to calculate the clinical similarity of small molecules.[40,43] The range of clinical similarity S_a^s of small molecules is between 0 and 1.

2.2.1.2. Chemical Similarity. DrugBank database[42] contains chemical structure information (SMILES format) and the Open Babel[44] can be utilized to compute MACCS fingerprints of each drug. Chemical similarity is widely used in drug discovery and drug combinations,[40,41] and its value ranges from 0 to 1.

Considering that there may be bias in the calculation of individual small molecule similarity, a weighted combination strategy is used to integrate two similarities as follows

$$S^s = \frac{(\mu_1 S_a^s + \mu_2 S_c^s)}{\sum_i \mu_i}; \ (i = 1, \ 2) \tag{1}$$

where S_a^s and S_c^s represent drug clinical similarity and chemical similarity, respectively. We set the parameter $\mu_i = 1$ $(i = 1, 2)$, and each similarity has the same weight.

2.2.2. Representation of miRNA Similarities. *2.2.2.1. Functional Similarity.* Accumulated studies show that the functional similarity of miRNAs can greatly improve the performance of small molecule−miRNA association prediction.[28] Thus, we use gene functional similarities and miRNA−gene associations to calculate miRNAs' similarity. Refer to the calculation steps of Xiao et al.[45] The functional similarity of miRNA m_i and m_j is calculated by the best-match average (BMA) method[46] as follows

$$S_f^m(m_i, m_j) = \frac{\sum_{g \in G_i} S(g, G_j) + \sum_{g \in G_j} S(g, G_i)}{|G_i| + |G_j|} \tag{2}$$

where G_i and G_j represent the gene sets associated with m_i and m_j, respectively. $|\cdot|$ denotes the number of genes in the sets.

2.2.2.2. Sequence Similarity. miRNA sequence similarity is widely used to improve the performance of miRNA-related association prediction because miRNA sequences contain rich biological information.[28] Meanwhile, we download the miRNA sequences from miRBase data set.[47] miRNAs are single-stranded small RNAs of approximately 21−23 bases (A, U, G, and C) in size. There are 16 types of base-pairs (AA, AU, AG, ..., CG, CC) through the two bases that are combined. Therefore, each miRNA sequence can be represented as a 16-dimensional vector, where the value is the frequency at which the corresponding base combination appears.[48] Finally, we use cosine similarity to calculate the feature similarity of miRNAs as follows

$$S_{\mathrm{s}}^{\mathrm{m}}(m_i, m_j) = \frac{M_i{*}M_j}{\| M_i \| \| M_i \|} \tag{3}$$

where M_i and M_j represent feature vectors of miRNA m_i and m_j, respectively.

Similar to the drug similarity calculation, considering that there may be bias in the calculation of individual miRNA similarity, a weighted combination strategy is used to integrate two similarities as follows

$$S^{\mathrm{m}} = \frac{(\tau_1 S_{\mathrm{f}}^{\mathrm{m}} + \tau_2 S_{\mathrm{s}}^{\mathrm{m}})}{\sum_i \tau_i}; \; (i = 1, 2) \tag{4}$$

where $S_{\mathrm{f}}^{\mathrm{m}}$ and $S_{\mathrm{s}}^{\mathrm{m}}$ represent miRNA functional similarity and feature similarity, respectively, and their value ranges from 0 to 1. We set the parameter $\tau_i = 1$ $(i = 1, 2)$, and each similarity has the same weight.

2.2.3. Representation of Disease Similarity. In this study, we can download the denominations of diseases, which can be represented as a directed acyclic graph (DAG)[49] from the MeSH database (http://www.ncbi.nlm.nih.gov/). In a DAG, a disease is represented by a node, and the relationship between diseases is represented by the link. A disease could be defined as $\mathrm{DAG}_d = (d, S_d, L_d)$, where S_d is the set of all nodes (including node d itself) and L_d is the set of links. The semantic contribution of a disease t to disease d can be obtained as follows

$$\begin{cases} D_d(t) = 1 \\ D_d(t) = \max\{\Delta{*}D_d(t') | t' \in \text{children of} t\} \text{ if } t \neq d \end{cases} \tag{5}$$

where Δ represents the semantic contribution parameter. According to the literature,[49] we set $\Delta = 0.5$. According to the definition of semantic contribution value, the semantic value of disease d can be obtained as follows

$$\mathrm{DV}(d) = \sum_{t \in S_d} D_d(t) \tag{6}$$

The more similar the DAG of the two diseases, the higher the similarity between the two diseases. We can compute the semantic similarity of disease d_i and d_j based on the following equation

$$S^{\mathrm{d}}(d_i, d_j) = \frac{\sum_{t \in S_{d_i} \cap S_{d_j}} (D_{d_i}(t) + D_{d_j}(t))}{\mathrm{DV}(d_i) + \mathrm{DV}(d_j)} \tag{7}$$

where $\mathrm{DV}(d_i)$ and $\mathrm{DV}(d_j)$ represent the semantic values of disease t related to disease d_i and d_j, respectively. The range of semantic similarity is between 0 and 1.

2.2.4. Representation of Heterogeneous Relationships. The small molecule−miRNA network is constructed by the known small molecule−miRNA association. Let matrix $Y_{\mathrm{sm}} \in \mathbb{R}^{N_{\mathrm{s}} \times N_{\mathrm{m}}}$ represent the association between N_{s} small molecules and N_{m} miRNAs. If small molecule s_i was observed to be associated with miRNA m_j, then $(Y_{\mathrm{sm}})_{ij}$ is 1, otherwise 0. The small molecule−miRNA association data set can be obtained from SM2miR.[50] Same representation as small molecule−disease associations, the matrix $Y_{\mathrm{sd}} \in \mathbb{R}^{N_{\mathrm{s}} \times N_{\mathrm{d}}}$ represents the association case between N_{s} small molecules and N_{d} diseases, and the matrix $Y_{\mathrm{md}} \in \mathbb{R}^{N_{\mathrm{m}} \times N_{\mathrm{d}}}$ represents the association between N_{m} miRNAs and N_{d} diseases. The small molecule−disease association data set and miRNA−disease association data set can be downloaded from the comparative toxicogenomics database (CTD)[51] and HMDD,[52] respectively.

2.3. *K* Nearest Neighbor Profiles (KNNPs).

To predict novel small molecule (miRNA), which has no associations with other miRNA (small molecule), the K nearest neighbor profiles can be utilized to solve this problem.[45] Assume that $\mathrm{set_s} = \{s_1, s_2, ..., s_{N_{\mathrm{s}}}\}$ and $\mathrm{set_m} = \{m_1, m_2, ..., m_{N_{\mathrm{m}}}\}$ represent the set of N_{s} small molecules and N_{m} miRNAs, respectively. In matrix Y_{sm}, the ith row vector $Y_{\mathrm{sm}}^{\mathrm{s}}(s_i) = (Y_{i1}, Y_{i2}, ..., Y_{iN_{\mathrm{m}}})$ indicates the interaction profile for small molecule s_i, and the jth column vector $Y_{\mathrm{sm}}^{\mathrm{m}}(m_j) = (Y_{j1}, Y_{j2}, ..., Y_{jN_{\mathrm{s}}})$ represents the interaction profile for miRNA m_j. For each small molecule s_{q}, we can obtain a novel interaction profile based on its similarities with other K nearest known small molecules as follows

$$Y_{\mathrm{sm}}^{\mathrm{s}}(s_{\mathrm{q}}) = \frac{1}{Q_{\mathrm{s}}} \sum_{i=1}^{K} \omega_i Y_{\mathrm{sm}}^{\mathrm{s}}(s_i) \tag{8}$$

where s_1 to s_K represent small molecules that are sorted in descending order based on their similarity to s_{q}. The weight parameter can be denoted as $\omega_i = \varepsilon^{i-1} \times S^{\mathrm{s}}(s_i, s_{\mathrm{q}})$, indicating that the more similar s_i is to s_{q}, the higher the weight that is assigned. $\varepsilon \in [0,1]$ is a decay coefficient, and $Q_{\mathrm{s}} = \sum_{1 \le j \le K} S^{\mathrm{s}}(s_i, s_{\mathrm{q}})$ represents the normalization term. Similarly, the novel interaction profile for each miRNA m_i can be defined as follows

$$Y_{\mathrm{sm}}^{\mathrm{m}}(m_{\mathrm{q}}) = \frac{1}{Q_{\mathrm{m}}} \sum_{j=1}^{K} \omega_j Y_{\mathrm{sm}}^{\mathrm{m}}(m_j) \tag{9}$$

where m_1 to m_K represent miRNAs that are sorted in descending order based on their similarity to m_{q}. ω_j denotes the weight parameter, and $Q_{\mathrm{m}} = \sum_{1 \le j \le K} S^{\mathrm{m}}(s_j, s_{\mathrm{q}})$ represents the normalization term. Afterward, the adjacency matrix Y can be updated as follows

$$Y_{\mathrm{sm}} = \max(Y_{\mathrm{sm}}, \overline{Y}_{\mathrm{sm}}) \tag{10}$$

where $\overline{Y}_{\mathrm{sm}} = (a_1 Y_{\mathrm{sm}}^{\mathrm{s}} + a_2 Y_{\mathrm{sm}}^{\mathrm{m}}) / \sum a_i$ $(i = 1, 2)$ and a_i is a weight parameter. According to experience, we assign the same weight to each separated similarity, that is, set $a_1 = a_2 = 1$. In the same manner, the novel adjacency matrices Y_{sd} and Y_{md} can be obtained based on K nearest neighbor profiles.

2.4. Small Molecule−miRNA Association Prediction Model.

2.4.1. Modeling Prior Knowledge of Similarities. In a heterogeneous network, the similarity between entities contains a variety of biological information. Therefore, we use three types of similarities to construct a prediction model. To build the model, we refer to Xuan's method.[20,53] First, let $C = (C_{ij}) \in \mathbb{R}^{N_{\mathrm{s}} \times N_{\mathrm{m}}}$ represent the association score matrix between small molecule and miRNA, where $C_{ij} \ge 0$ indicates the probability score associated with small molecule s_i and miRNA m_j. The ith row of matrix C is denoted as C_i representing the likelihood of association between small molecule i and all miRNAs, and the jth column of matrix C is denoted as $(C^{\mathrm{T}})_j$, representing the likelihood that miRNA j is associated with all small molecules. The more two small molecules are associated with similar miRNAs, the more similar they are. Therefore, (C_i) $(C^{\mathrm{T}})_j = (CC^{\mathrm{T}})_{ij}$ can represent the similarity of small molecule s_i and s_j, and $(S^{\mathrm{s}})_{ij}$ indicates the known clinical similarity between small molecule s_i and small molecule s_j. We use the clinical similarity to constrain the update of matrix C for incorporating small molecule clinical similarity into matrix C. Then, small molecule similarity matrix S^{s} can be factorized as CC^{T} as follows

$$\min_{C\geq 0} \| S^{s} - CC^{T} \|_{F}^{2} \tag{11}$$

where a$\|\cdot$a$\|_{F}$ represents the Forbenius norm of a matrix. For entity miRNAs and diseases, they have the same properties. We can make full use of the miRNA similarity and disease similarity and use the same factorization method to calculate the association score matrix based on the corresponding actual similarity. Therefore, we can obtain the formula

$$\begin{aligned}\min_{A\geq 0,B\geq 0,C\geq 0} & \| S^{s} - CC^{T} \|_{F}^{2} + \| \gamma(S^{m} - C^{T}C \|_{F}^{2} \\ & + \| S^{m} - BB^{T} \|_{F}^{2} + \| S^{s} - AA^{T} \|_{F}^{2} + \| S^{d} - A^{T}A \|_{F}^{2} \\ & + \| S^{d} - B^{T}B \|_{F}^{2}) \end{aligned} \tag{12}$$

where $A = (A_{ij}) \in \mathbb{R}^{N_{s} \times N_{d}}$ represents the association score matrix between small molecule and disease, $B = (B_{ij}) \in \mathbb{R}^{N_{m} \times N_{d}}$ represents the association score matrix between miRNA and disease, and S^{m} and S^{d} represent the known miRNA similarity and disease similarity, respectively. γ is a hyperparameter that adjusts the contribution of prior knowledge.

2.4.2. Modeling the Node Associations. After the KNNP operation, the association scores of small molecule−miRNA association matrix $Y_{sm} \in \mathbb{R}^{N_{s} \times N_{m}}$, small molecule−disease association matrix $Y_{sd} \in \mathbb{R}^{N_{s} \times N_{d}}$, and miRNA−disease association matrix $Y_{md} \in \mathbb{R}^{N_{m} \times N_{d}}$ are between 0 and 1. The number of nonzero elements in matrices Y_{sm}, Y_{sd}, and Y_{md} are much less than zero elements so that the process of model optimization is based on known relationships. $Y_{A} \in \mathbb{R}^{N_{s} \times N_{d}}$ is an indicator matrix; if (s_{i},d_{j}) is known small molecule−disease associations, then $(Y_{A})_{ij} = 1$, otherwise 0. In a similar way, we can obtain $Y_{B} \in \mathbb{R}^{N_{m} \times N_{d}}$ and $Y_{C} \in \mathbb{R}^{N_{s} \times N_{m}}$. Obviously, Y_{A}, Y_{B}, and Y_{C} are, respectively, equal to Y_{sd}, Y_{md}, and Y_{sm} before the KNNP operation. Note that A, B, and C represent the association score matrix of small molecule−disease, miRNA−disease, and small molecule−miRNA, respectively, and the association score matrix should be close to the known association matrix; therefore, the constraint can be obtained as follows

$$\begin{aligned}\min_{A\geq 0,B\geq 0,C\geq 0} & \alpha(\| Y_{A} \odot (Y_{sd} - A)\|_{F}^{2} + \| Y_{B} \odot (Y_{md} - B)\|_{F}^{2} \\ & + \| Y_{C} \odot (Y_{sm} - C)\|_{F}^{2}) \end{aligned} \tag{13}$$

where α is a hyperparameter that adjusts the contribution of the node associations and $\odot$ is the Hadamard product.

2.4.3. Modeling the Topology Characteristics of Associations. In a heterogeneous network consisting of small molecules, miRNAs, and diseases, it is well known that the more the same diseases associated with small molecule s_{i} and miRNA m_{j}, the more likely the small molecule s_{i} is related to miRNA m_{j}. Since A_{i} indicates the likelihood of association of small molecule s_{i} with all diseases and $(B^{T})_{j}$ indicates the likelihood of association of miRNA m_{j} with all diseases, $(AB^{T})_{ij}$ can represent the likelihood of association between small molecule s_{i} and miRNA m_{j}. However, C_{ij} also represents the likelihood of association between small molecule s_{i} and miRNA m_{j}; therefore, we can minimize the interpolation of AB^{T} and C to obtain the constraints of A, B, and C for the association score matrix tends to be more realistic. The specific formula is as follows

$$\min_{A\geq 0,B\geq 0,C\geq 0} \beta \| C - AB^{T} \|_{F}^{2} \tag{14}$$

where β is a hyperparameter that adjusts the contribution of the topology characteristics.

2.4.4. Considering the Sparseness of Associations. It is well known that the potential associations of small molecule−miRNA, small molecule−disease, and miRNA−disease are sparse.[45] Therefore, the l_{1}-regularization is applied to matrices A, B, and C simultaneously for studying sparse underlying associations. Meanwhile, the use of l_{1}-regularization can prevent the model from overfitting. The objective function can be obtained as follows

$$\min_{A\geq 0,B\geq 0,C\geq 0} \delta(\| A \|_{1} + \| B \|_{1} + \| C \|_{1}) \tag{15}$$

where δ is a hyperparameter that controls the contribution of the sparse term.

The objective functions of modeling the prior knowledge of similarities in eq 12, modeling the node associations in eq 13, modeling the topology characteristics of associations in eq 14, and modeling the sparseness of associations in eq 15 are combined into a unified objective function as follows

$$\begin{aligned}\min_{A\geq 0,B\geq 0,C\geq 0} & \| S^{s} - CC^{T} \|_{F}^{2} + \gamma \|(S^{m} - C^{T}C \|_{F}^{2} \\ & + \| S^{m} - BB^{T} \|_{F}^{2} + \| S^{s} - AA^{T} \|_{F}^{2} + \| S^{d} - A^{T}A \|_{F}^{2} \\ & + \| S^{d} - B^{T}B \|_{F}^{2}) + \alpha \|(Y_{A} \odot (Y_{sd} - A)\|_{F}^{2} \\ & + \| Y_{B} \odot (Y_{md} - B)\|_{F}^{2} + \| Y_{C} \odot (Y_{sm} - C)\|_{F}^{2}) + \beta \\ & \| C - AB^{T} \|_{F}^{2} + \delta(\| A \|_{1} + \| B \|_{1} + \| C \|_{1}) \end{aligned} \tag{16}$$

2.5. Initialization of the Association Score Matrices.

This model relies on the selection of initial values in the iterative process, and the initial values are not unique. Initial values of A, B, and C directly affect the quality of the results. In the SMMART model, we use an improved singular value factorization method[54] to initialize A, B, and C. We decomposed $Y_{sm} \in \mathbb{R}^{N_{s} \times N_{m}}$ into $U_{sm} \in \mathbb{R}^{N_{s} \times N_{s}}$, $\sum_{sm} \in \mathbb{R}^{N_{s} \times N_{m}}$, and $V_{sm} \in \mathbb{R}^{N_{m} \times N_{m}}$, where $\sum_{sm}$ is a diagonal matrix and the value of the diagonal is a singular value. After normalizing matrices U_{sm} and V_{sm}, we obtain the initialized $C = U_{sm}\sum_{sm}V_{sm}^{T}$. In the same way, we can obtain A and B after initialization.

2.6. Optimization.

It is complicated to solve the objective function eq 16 directly; thus, we decompose the optimization problem into several subproblems and then optimize the subproblems iteratively.

2.6.1. C-Subproblem. We use the method of controlling variables,[55] that is, the values of A and B are fixed when updating C. The subproblem can be solved as follows

$$\begin{aligned}\min_{C\geq 0} L(C) = & \| S^{s} - CC^{T} \|_{F}^{2} + \| \gamma S^{m} - C^{T}C \|_{F}^{2} \\ & + \alpha \| Y_{C} \odot (Y_{sm} - C)\|_{F}^{2} + \beta \\ & \| C - AB^{T} \|_{F}^{2} + \delta \| C \|_{1} \end{aligned} \tag{17}$$

By setting the derivative of $L(C)$ with respect to C to 0, the formula can be obtained as

$$
\begin{aligned}
L(C) &= -4(S^{\mathrm{s}} - CC^{\mathrm{T}})C - 4\gamma C(S^{\mathrm{m}} - C^{\mathrm{T}}C) \\
&\quad + 2\alpha(Y_C \odot (C - Y_{\mathrm{sm}})) + 2\beta C - 2\beta AB^{\mathrm{T}} + \delta O_C \\
&= 0;\ C \geq 0
\end{aligned} \tag{18}
$$

where $O_C \in \mathbb{R}^{N_s \times N_m}$ is an identity matrix with all elements being 1. According to the Lagrange multipliers,[56] by multiplying both sides of eq 16 by C_{ij}, we obtain

$$
\begin{aligned}
&(-4S^{\mathrm{s}}C + 4CC^{\mathrm{T}}C - 4\gamma CS^{\mathrm{m}} + 4\gamma CC^{\mathrm{T}}C \qquad ; C_{ij} = 0 \\
&+ 2\alpha Y_C \odot C - 2\alpha Y_C \odot Y_{\mathrm{sm}} + 2\beta C - 2\beta AB^{\mathrm{T}} \\
&+ \delta O_C)_{ij}
\end{aligned} \tag{19}
$$

Therefore, we determine the update rules as follows

$$
C_{ij}^{t+1} \leftarrow C_{ij}^{t} \cdot \frac{(4S^{\mathrm{s}}C + 4\gamma CS^{\mathrm{m}} + 2\alpha Y_C \odot Y_{\mathrm{sm}} + 2\beta AB^{\mathrm{T}})_{ij}}{(4CC^{\mathrm{T}}C + 4\gamma CC^{\mathrm{T}}C + 2\alpha Y_C \odot C + 2\beta C + \delta O_C)_{ij}} \tag{20}
$$

The matrix C is updated based on eq 18 until convergence.

2.6.2. B-Subproblem. When updating B, the values of C and A are fixed. The subproblem can be solved as follows

$$
\begin{aligned}
\min_{B \geq 0} L(B) &= \gamma \| S^{\mathrm{m}} - BB^{\mathrm{T}} \|_{\mathrm{F}}^2 + \gamma \| S^{\mathrm{d}} - B^{\mathrm{T}}B \|_{\mathrm{F}}^2 \\
&\quad + \alpha \| Y_B \odot (Y_{\mathrm{md}} - B) \|_{\mathrm{F}}^2 + \beta \| C - AB^{\mathrm{T}} \|_{\mathrm{F}}^2 \\
&\quad + \delta \| B \|_1
\end{aligned} \tag{21}
$$

By setting the derivative of $L(B)$ with respect to B to 0, the formula can be obtained as follows

$$
\begin{aligned}
L(B) &= -4\gamma(S^{\mathrm{m}} - BB^{\mathrm{T}})B - 4\gamma E(S^{\mathrm{d}} - B^{\mathrm{T}}B) \\
&\quad + 2\alpha(Y_{\mathrm{B}} \odot (B - Y_{\mathrm{md}})) - 2\beta C^{\mathrm{T}}A + 2\beta BA^{\mathrm{T}}A + \delta O_B \\
&= 0;\ B \geq 0
\end{aligned} \tag{22}
$$

where $O_B \in \mathbb{R}^{N_m \times N_d}$ is an identity matrix with all elements being 1. According to the Lagrange multipliers,[56] by multiplying both sides of eq 20 by B_{ij}, we obtain

$$
\begin{aligned}
&(-4\gamma S^{\mathrm{m}}B + 4\gamma BB^{\mathrm{T}}B - 4\gamma BS^{\mathrm{d}} + 4\gamma BB^{\mathrm{T}}B \qquad ; B_{ij} = 0 \\
&+ 2\alpha Y_{\mathrm{B}} \odot B - 2\alpha Y_{\mathrm{B}} \odot Y_{\mathrm{md}} - 2\beta C^{\mathrm{T}}A + 2\beta BA^{\mathrm{T}}A \\
&+ O_B)_{ij}
\end{aligned} \tag{23}
$$

Therefore, we determine the update rules as follows

$$
B_{ij}^{t+1} \leftarrow B_{ij}^{t} \cdot \frac{(4\gamma S^{\mathrm{m}}B + 4\gamma BS^{\mathrm{d}} + 2\alpha Y_{\mathrm{B}} \odot Y_{\mathrm{md}} + 2\beta C^{\mathrm{T}}A)_{ij}}{(8\gamma BB^{\mathrm{T}}B + 2\alpha Y_{\mathrm{B}} \odot B + 2\beta BA^{\mathrm{T}}A + \delta O_{\mathrm{B}})_{ij}} \tag{24}
$$

The matrix B is updated based on eq 22 until convergence.

2.6.3. A-Subproblem. When updating A, the values of B and C are fixed. The subproblem can be solved as follows

$$
\begin{aligned}
\min_{A \geq 0} L(A) &= \gamma \| S^{\mathrm{s}} - AA^{\mathrm{T}} \|_{\mathrm{F}}^2 + \gamma \| S^{\mathrm{d}} - A^{\mathrm{T}}A \|_{\mathrm{F}}^2 \\
&\quad + \alpha \| Y_A \odot (Y_{\mathrm{sd}} - A) \|_{\mathrm{F}}^2 + \beta \| C - AB^{\mathrm{T}} \|_{\mathrm{F}}^2 \\
&\quad + \delta \| A \|_1
\end{aligned} \tag{25}
$$

By setting the derivative of $L(B)$ with respect to B to 0, the formula can be obtained as follows

$$
\begin{aligned}
L(A) &= -4\gamma(S^{\mathrm{s}} - AA^{\mathrm{T}})A - 4\gamma A(S^{\mathrm{d}} - A^{\mathrm{T}}A) \\
&\quad + 2\alpha(Y_A \odot (A - Y_{\mathrm{sd}})) - 2\beta CB + 2\beta AB^{\mathrm{T}}B + \delta O_A \\
&= 0;\ A \geq 0
\end{aligned} \tag{26}
$$

where $O_A \in \mathbb{R}^{N_s \times N_d}$ is an identity matrix with all elements being 1. According to the Lagrange multipliers,[56] by multiplying both sides of eq 24 by A_{ij}, we obtain

$$
\begin{aligned}
&(-4\gamma S^{\mathrm{s}}A + 4\gamma AA^{\mathrm{T}}A - 4\gamma AS^{\mathrm{d}} + 4\gamma AA^{\mathrm{T}}A \quad A_{ij} = 0 \\
&+ 2\alpha Y_A \odot A - 2\alpha Y_A \odot Y_{\mathrm{sd}} - 2\beta CB \\
&+ 2\beta AB^{\mathrm{T}}B + \delta O_A)_{ij}
\end{aligned} \tag{27}
$$

Therefore, we determine the update rules as follows

$$
A_{ij}^{t+1} \leftarrow A_{ij}^{t} \cdot \frac{(4\gamma S^{\mathrm{s}}A + 4\gamma AS^{\mathrm{d}} + 2\alpha Y_A \odot Y_{\mathrm{sd}} + 2\beta CB)_{ij}}{(8\gamma AA^{\mathrm{T}}A + 2\alpha Y_A \odot A + 2\beta AB^{\mathrm{T}}B + \delta O_A)_{ij}} \tag{28}
$$

The matrix A is updated based on eq 26 until convergence.

After obtaining the final matrices A, B, and C, we can obtain the predicted correlation scores of small molecules and miRNAs from matrix C and obtain the predicted correlation scores of small molecules and diseases from matrix B. In the next section, the entire experimental process is described in detail. Overall, the complexity of our method is $O(KN_s + KN_m + {}_{\mathrm{Niter}})$, where K is the number of the nearest neighbor profiles, N_s and N_m represent the number of small molecules and miRNAs, respectively, and ${}_{\mathrm{Niter}}$ is optimization iteration times. We run the codes at 2.6 GHz Intel(R) Core(TM) i7-9750H CPU with 32 GB RAM, and the running time of SMMART is 51 s.

2.7. Software Package. We upload the R software package through GitHub to https://github.com/CS-BIO/SMMART, containing all data sets and codes. Furthermore, the package can be used to execute 5-fold cross-validation, as well as select hyperparameters for reproducing the results.

3. RESULTS

3.1. Data Collection and Preprocessing. The gold standard data set for discovering potential small molecule−miRNA associations are obtained from SM2miR,[50] which is a database that contains a total of 5112 experimentally verified associations. First, different small molecule−miRNA pairs with the same mature miRNA can be merged. Then, the same associations need to be removed. Therefore, all miRNA calculations involved in this model are based on the precursor miRNA. Finally, a total of 4182 experimentally verified associations can be obtained from SM2miR,[50] which includes 251 small molecules and 901 miRNAs. The comparative toxicogenomics database (CTD)[51] provides association data between small molecules and diseases. We download miRNA−disease association data from the HMDD database.[52] To construct a heterogeneous network, the numbers of small molecules, miRNAs, and diseases need to be unified. The numbers of small molecules and miRNAs are based on SM2miR[50] (including 251 small molecules and 901 miRNAs). To prevent the association data from being too sparse, we choose the intersection of diseases in the CTD database[51] and the HMDD database[52] (obtaining 361 diseases). The chemical

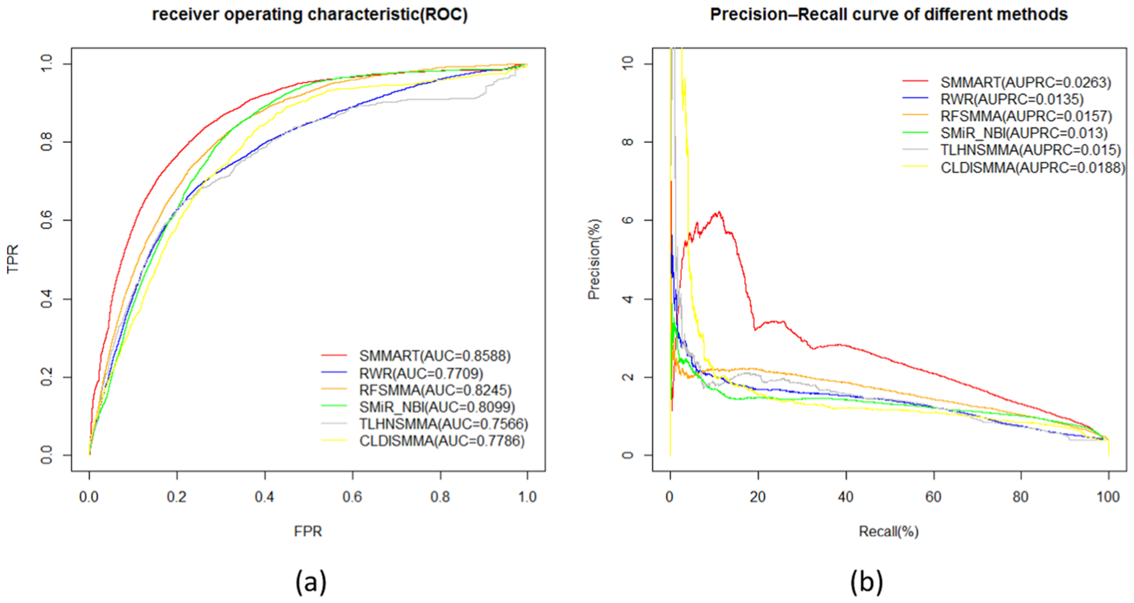


Figure 2. (a) ROC and AUC of six methods on small molecule−miRNA association prediction task and (b) precision/recall curve and AUPRC of six methods on small molecule−miRNA association prediction task.

Table 1. *P*-Values Obtained through a Paired *t*-Test of the AUCs and AUPRCs of SMMART and Other Compared Methods for 10 Runs

	P-value				
	RWR	RFSMMA	SMiR_NBI	TLHNSMMA	CLDISMMA
AUCs	1.667×10^{-11}	$2.92E \times 10^{-8}$	5.207×10^{-10}	1.023×10^{-9}	1.168×10^{-11}
AUPRCs	7.496×10^{-10}	5.854×10^{-10}	2.644×10^{-10}	1.913×10^{-7}	4.38×10^{-6}

structure and the ATC code of the small molecule for calculating chemical similarity and clinical similarity are from the DrugBank database.[42] The miRTarbase database[57] and miRbase database[47] provide miRNA−gene association data and sequence data for computing miRNA functional similarity and sequence similarity. We download disease semantic data from the Mesh database to calculate disease semantic similarity. Finally, small molecule similarity matrix $S^s \in \mathbb{R}^{251 \times 251}$, miRNA similarity matrix $S^m \in \mathbb{R}^{901 \times 901}$, and disease similarity matrix $S^d \in \mathbb{R}^{361 \times 361}$ can be obtained for discovering unknown associations between small molecules and miRNAs. The detailed representation of all data is shown in Table S1.

3.2. Experimental Settings. To study the performance of SMMART, 5-fold cross-validation is applied and repeated 10 times. For each data set, we randomly divide the associations between small molecules and miRNAs into five parts of the same size. Each part takes turns as the test set and the remaining four parts as training sets. We performed a total of 10 times of 5-fold cross-validation and averaged each group of TPR, FPR, recall, and precision obtained for each 5-fold. The final average value was used to calculate AUC, AUPR value, draw receiver operating characteristic (ROC) curve, and precision−recall (PR) curve. For SMMART method, the regularization coefficient α is selected from $\{10^{-3}, 10^{-2}, 10^{-1}, 10^{0}, 10^{1}\}$, γ is chosen from {1, 5, 10, 15, 20}, δ is obtained from {1, 10, 50, 100, 150} and β is acquired from {0.1, 0.5, 1, 5, 10}. The k of K nearest neighbor profiles is chosen from {1, 2, 3, 4, 5}.

We consider several evaluation metrics to evaluate the performance of association prediction results between small molecules and miRNAs. The recall, specificity, G_mean, and precision are obtained as follows: recall = TP/(TP + FN), specificity = TN/(TN + FP), G_mean = $(\text{recall} \times \text{specificity})^{0.5}$, and precision = TP/(TP + FP). TP, FP, TN, and FN represent true positive, false positive, true negative, and false negative, respectively. We also use the area under the ROC (AUC[58]), the area under the precision/recall curve (AUPRC), and P-value (calculated by a paired t-test[59]) as important evaluation metrics to evaluate the overall performance.

3.3. Cross-Validation Experiments. We compare the SMMART model with several state-of-the-art models (see Section 1 for more details), namely: RWR,[29] RFSMMA,[28] SMiR_NBI,[30] TLHNSMMA,[36] and CLDISMMA.[37] The data sets of the SMMART model are used for the comparison method simultaneously. The selection of hyperparameters for the comparison method is detailed in the Supporting Information. The AUCs of SMMART, RWR, RFSMMA, SMiR_NBI, TLHNSMMA, and CLDISMMA are 0.8588, 0.7709, 0.8245, 0.8099, 0.7566, and 0.7786, respectively, where the AUC of SMMART is significantly higher than those of other methods (Figure 2a). Figure 2b shows that SMMART achieves better performance than the other five methods: RWR (AUPRC = 0.0135), RFSMMA (AUPRC = 0.0157), SMiR_NBI (AUPRC = 0.013), TLHNSMMA (AUPRC = 0.015), and CLDISMMA (AUPRC = 0.0188). Furthermore, the AUCs and AUPRCs of SMMART and the other methods with different runs were compared using a paired t-test via 5-fold cross-validation. As shown in Table 1, the p-values were less than 0.05, suggesting that the differences between AUCs and AUPRCs were statistically significant. To further verify the performance of the SMMART model, we

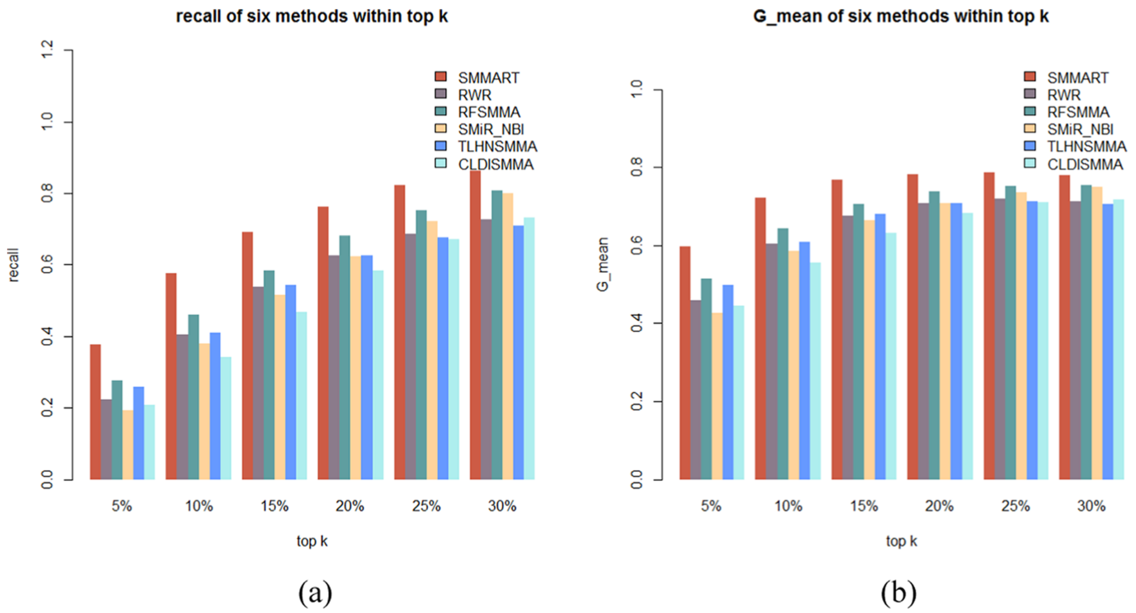


Figure 3. (a) Recall of six tested methods on small molecule−miRNA association prediction task and (b) *G*_mean of six tested methods on small molecule−miRNA association prediction task.

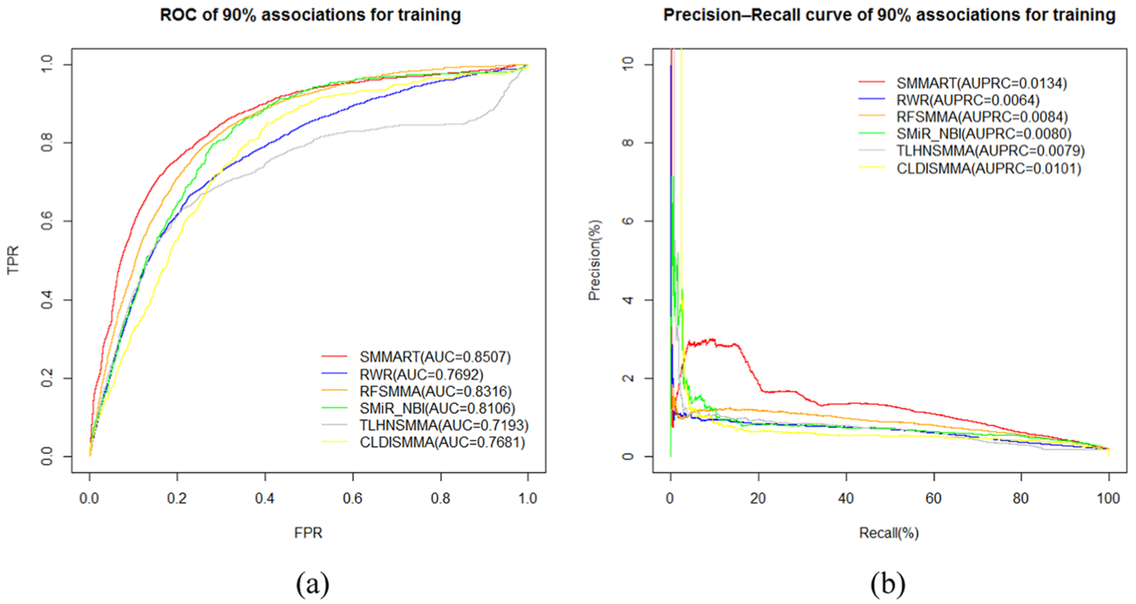
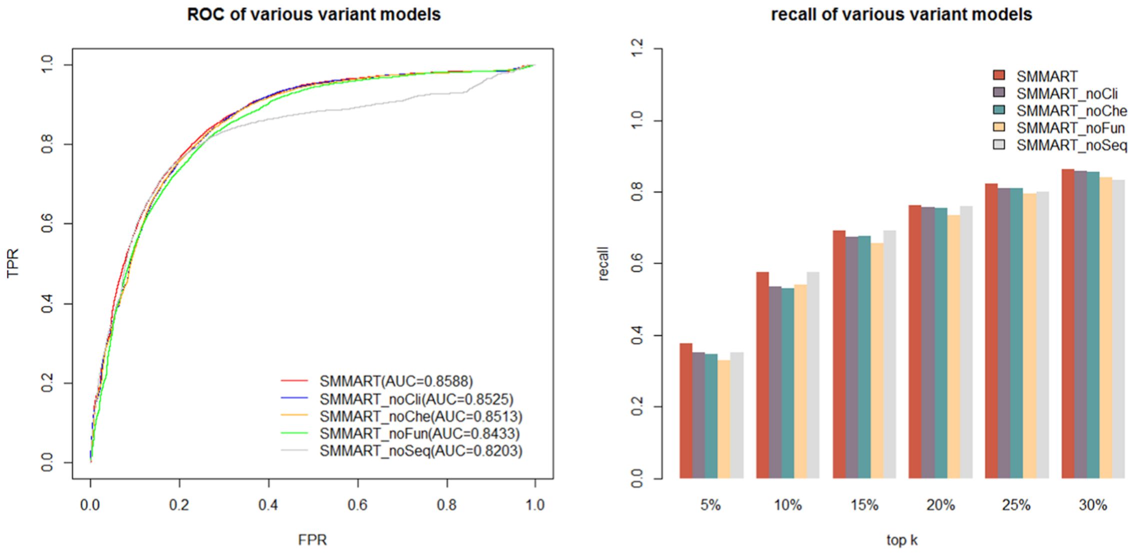


Figure 4. Comparison of SMMART with the other methods based on 90% known associations. (a) ROC and AUC of six methods on small molecule−miRNA association prediction task and (b) precision/recall curve and AUPRC of six methods on small molecule−miRNA association prediction task.

compare recall and *G*_mean with the other five comparison methods. Figure 3a,b shows the recall and *G*_mean of the six methods, which indicate that SMMART achieves higher performance than the other methods within top 5−30%. Table S2 contains detailed numerical comparison information of various indicators. For example, precision of the SMMART model is better than that of any comparison method. There are two main reasons why the SMMART model can achieve good results under evaluation metrics. First, the SMMART model effectively integrates small molecule−miRNA associations and small molecule−disease associations at a system level. Second, the graph regularization technology in the SMMART model makes the model achieve better performance.

3.4. Parameter Sensitivity Analysis.

In this section, we investigate the parameter results of SMMART. There are four regularization parameters in SMMART, i.e., α, γ, δ, and β. More specifically, α trades off the contribution of the association relationships, γ adjusts the contribution of the corresponding prior knowledge, δ trades off the contribution of the sparse term, and β adjusts the contribution of the topology characteristics. There is a hyperparameter *k* from *K* nearest neighbor profiles, and the parameter *k* adjusts the number of most similar neighbors. In this study, we first analyze the regularization parameter sensitivity. We fix three of the four parameters and tune the other one from the candidate set. For four regularization parameters, first, we tune α from $\{10^{-3}, 10^{-2}, 10^{-1}, 10^{-0}, 10\}$ by fixing $\gamma = 10$, $\delta = 10$, and $\beta = 1$. Figure



(a) (b)

Figure 5. Performance comparison between SMMART and variant models. (a) AUC of SMMART and variant models. (b) Recall of SMMART and variant models within top 5−30%.

S1a shows the AUC of SMMART by tuning α, and the best result can be obtained when α = 0.01. Then, we fix α = 0.01, δ = 10, and β = 1 to tune γ from {1, 5, 10, 15, 20}. Figure S1b indicates the AUC of our method by tuning γ, and we can acquire the highest AUC values 0.8582 when γ = 5. Next, we tune δ from {5, 10, 50, 100, 150} by fixing α = 0.01, γ = 5, and β = 1 in the same way. The results of tuning δ are shown in Figure S1c, and we can get the highest AUC score when δ = 100. Finally, we fix α = 0.01, γ = 5, and δ = 100 to tune β from {0.1, 0.5, 1, 5, 10}, and the highest AUC value is 0.8588 when β = 5, which is shown in Figure S1d. The best regularization parameter combination we can obtain is α = 0.01, γ = 5, δ = 100, and β = 5. Based on the best combination of regularization parameters, we analyzed the k value in K nearest neighbor profiles. As shown in Figure S1e, when $k \leq 3$, the AUC shows a rising trend as the value of k increases, probably because of the introduction of the node neighbor information. When $k > 3$, the AUC value decreases slowly, probably because some noise information is introduced, which has a negative impact on the result. When k = 1, it does not consider the result of the K nearest neighbor. Although the effect is not very large, the prediction of the associations between new miRNAs has a greater impact.

3.5. Robustness Analysis. To analyze the performance of the SMMART model on different sparse data sets, the different proportions (90, 80, 70, 60, and 50%) of the training sets are used to experiment separately. As the variance of the data set can be quite high, we repeat the process 10 times and report the averaged AUC, AUPRC, recall, specificity, G_mean, and precision. Figure 4 shows the performance of SMMART and other methods by using 90% association data as the training set. As shown in Figure 4a, the AUC of the six methods is 0.8507, 0.7692, 0.8316, 0.8106, 0.7193, and 0.7681, respectively, which indicates that the SMMART model obtains the best result. The precision/recall curve of the six methods are shown in Figure 4b, and the SMMART model again achieves better performance. At the same time, we compare the recall, specificity, G_mean, and precision of the six methods within top k. As shown in Table S3, the recall of SMMART can reach 0.8474 in the top 30%, which is 0.12, 0.02, 0.04, 0.15, and 0.12 higher than the other five methods, respectively. The specificity, G_mean, and precision are also significantly higher than those of the other five methods. Figures S2−S5 and Tables S4−S7 show the performance of our model and other models when setting 80, 70, 60, and 50% as the training sets. Regardless of the proportion of the training set, SMMART always shows better performance. This result shows that SMMART is more capable of responding to changes in the data set and is more robust. There may be two main reasons for the SMMART model to show better performance even under unfamiliar sparse data sets: (1) a large heterogeneous information network containing three types of nodes is used to integrate multisource data, which provides a data basis for the good robustness of the model and (2) the graph regularization technology fully considers each type of data and provides technical support for the good robustness of the model.

3.6. Importance of KNNP. To overcome the characteristics of the sparseness of the association data and consider the association prediction of isolated points, the original adjacency matrices are processed by the K-nearest neighbor profiles (KNNPs). Figure S6 shows a comparison of pre- and non-pretreatment models, and the non-pretreatment model is called SMMART*. In Figure S6a,b, the recall and G mean are utilized as evaluation indicators to evaluate the prediction results of small molecule−miRNA associations. As shown in Figure S6a, recall of SMMART is 76% for the top 20% predicted candidates, significantly outperforming that of SMMART*(74%). Figure S6b shows that G_mean of SMMART is 77% for the top 15% predicted associations, also significantly outperforming that of SMMART*(75%).

3.7. Importance of Various Similarities. To illustrate that multiknowledge used to construct a similarity matrix is good for improving the performance, we used experiments to confirm the necessity of fusing each similarity data. In experiments, we removed the clinical similarity of small molecules (named SMMART_noCli), the chemical structure similarity of small molecules (named SMMART_noChe), the functional similarity of small molecules (named SMMART_-

Table 2. Top 15 Potential miRNA Candidates Discovered by SMMART for the Three Selected Small Molecules

rank	SM	miRNA	evidence	SM	miRNA	evidence	SM	miRNA	evidence
1	CD3385	mir-30b	25 526 515	CID 5793	mir-487a	unconfirmed	CID 119307	mir-27a	30 159 408
2	CD3385	let-7i	unconfirmed	CID 5793	mir-500a	unconfirmed	CID 119307	let-7g	unconfirmed
3	CD3385	mir-20b	23 617 628	CID 5793	mir-331	22 908 221	CID 119307	mir-181d	unconfirmed
4	CD3385	let-7c	25 951 903	CID 5793	mir-22	28 314 781	CID 119307	mir-195	22 893 786
5	CD3385	mir-449a	unconfirmed	CID 5793	mir-192	29 717 107	CID 119307	mir-497	30 108 441
6	CD3385	mir-98	25 526 515	CID 5793	mir-555	unconfirmed	CID 119307	let-7b	unconfirmed
7	CD3385	mir-320b	unconfirmed	CID 5793	mir-302d	unconfirmed	CID 119307	mir-196a	unconfirmed
8	CD3385	mir-107	26 636 340	CID 5793	mir-526a	unconfirmed	CID 119307	mir-744	30 159 408
9	CD3385	mir-29c	31 037 126	CID 5793	mir-518e	29 193 463	CID 119307	mir-383	24 555 688
10	CD3385	mir-200a	28 496 200	CID 5793	mir-520c	unconfirmed	CID 119307	mir-660	unconfirmed
11	CD3385	mir-455	unconfirmed	CID 5793	mir-520d	29 322 778	CID 119307	mir-202	30 809 600
12	CD3385	mir-133b	28 881 788	CID 5793	mir-103a	30 864 677	CID 119307	mir-505	unconfirmed
13	CD3385	mir-221	25 544 773	CID 5793	mir-557	unconfirmed	CID 119307	mir-33a	unconfirmed
14	CD3385	mir-26a	29 719 405	CID 5793	mir-26a	25 961 460	CID 119307	mir-335	unconfirmed
15	CD3385	mir-503	unconfirmed	CID 5793	mir-520h	unconfirmed	CID 119307	mir-433	unconfirmed

noFun), and the sequence similarity of small molecules (named SMMART_noSeq). Then, these four variant models are compared with the SMMART model, and the experimental results are shown in Figure 5. As shown in Figure 5a,b, on removing any similarity data, the performance of the model degrades. It is worth noting that the AUC and recall of the model SMMART_noCli are higher than those of the model SMMART_noChe, and those of the model SMMART_noFun are higher than those of the model SMMART_noSeq. These results show that the structure information of small molecules is a more important contribution to improve the performance of the model than clinical information. In a similar way, the sequence information of miRNAs is more helpful in improving the performance of the model than functional information.

3.8. Identification of Small Molecule−Disease Associations. This study takes into account the topological information in heterogeneous networks to discover unknown associations between small molecule and miRNA by correlating predictive small molecule−miRNA associations with small molecule−disease associations at a system level. The comparative experiments in Section 3.3 show that the SMMART model has great advantages in predicting the small molecule−miRNA associations. At the same time, we analyze the accuracy of the SMMART model in predicting small molecule−disease associations. The experimental results are shown in Tables S8 and S9. In addition to predicting the small molecule−miRNA associations, predicting the associations between small molecules and diseases also uses a 5-fold cross-validation. The experimental result is that the AUC value is 0.9422, and the top 20% recall reaches 98.99%. The detailed results are shown in Table S8. Table S9 is a literature search to verify the results of the 10 most relevant diseases predicted by the small molecule drug 5-fluorouracil (5-FU) (CID 3385) and glucose (CID 5793) through the SMMART model. For 5-fluorouracil, 8 of the 10 pairs with the highest correlation scores are validated. As for glucose, there are also six kinds of proven pairs. This shows that SMMART can also obtain good results in predicting the relationships between small molecule drugs and diseases, which is another advantage of the SMMART model.

3.9. Case Studies. To further analyze the prediction performance of SMMART, a case study is conducted for three types of small molecules, which include 5-fluorouracil (CD3385), glucose (CID 5793), and ginsenoside Rh2 (CID 119307). These three small molecules are most closely related to human life and health. 5-Fluorouracil plays a key role in the treatment of colon cancer. Meanwhile, 5-FU has serious cardiac toxicity that is displayed as cardiogenic shock, ventricular fibrillation, and myocardial infarction.[60] Glucose is essential for life, and a severe drop in blood glucose level can rapidly lead to coma and death.[61] Hexahydro-1,3,5-trinitro-1,3,5-triazine (RDX) has been classified as a class C potential human carcinogen by the U.S. Environmental Protection Agency.[62] We verify the predicted small molecule−miRNA associations by finding the literature in Pubmed. In this section, all of the known small molecule−miRNA associations are used to predict the associations. After being processed by the SMMART model, the association score of unknown small molecule−miRNA associations can be obtained. We take 15 miRNAs with the highest scores associated with the three small molecules and looked up the relationship pairs verified in the literature. The results are shown in Table 2, and 10, 7, and 6 of 15 candidate miRNAs are verified to be associated with 5-fluorouracil, glucose, and ginsenoside Rh2 by other literature. Taking small molecule drug 5-fluorouracil as an example, the literature[63] describes that the potential of mir-29c as a novel prognostic, treatment-predictive marker and diagnostic in ESCC and its therapeutic implication and mechanisms in overcoming 5-fluorouracil chemoresistance are explored. This expression indicates a resistance relationship between mir-29c and 5-FU. Similarly, the literature[64] describes that 5-fluorouracil and pirarubicin treatment can significantly induce the expression levels of miR-205 and miR-221, which fully explains the direct association between 5-fluorouracil and miR-221. Overall, the results of the case study further illustrate the accuracy of the SMMART model for predicting the association of small molecules with miRNAs.

4. CONCLUSIONS

In this study, we have developed a framework with graph regularization techniques that captures inter- and intrarelationships among small molecules, miRNAs, and diseases, aiming to infer unknown small molecule−miRNA associations. We first consider the sparseness of the association data, and the KNNP method is used to preprocess the associated data. On this basis, due to the existence of diverse information in heterogeneous networks, SMMART integrates prior knowledge, association information, and topological characteristics into the model through graph regularization techniques. Finally, we have validated the performance of SMMART through 5-fold cross-validation, robustness analysis, and case studies. Experimental results have shown that our model obtained great performance in inferring potential small molecule−miRNA associations.

We acknowledge that there are some limitations of SMMART under the current graph regularization technique-based prediction framework. First, the selection of an optimal parameter combination is a nontrivial work, and the model optimization process is complicated. Second, it is considered by assembling experimentally reported, large-scale data from publicly available databases in an overall framework. If the associations between small molecules and miRNAs can be predicted for specific cancer cell lines, better results may be achieved. Third, although the SMMART model has fused diverse information, the information is mainly network topology information. If more biological feature data (including more pharmacological information and published literature data, etc.) can be integrated, it may be of great help in improving the accuracy of the model. In future studies, incorporating and collecting more relevant complex network information through effective fusion methods (e.g., network embedding) from more literature and the databases may improve prediction performance. In addition, we will consider more forms of similarity calculation methods, including sequence alignment and Hamming distance of mature miRNAs and chemical structure similarity of small molecules. Furthermore, we may choose a better method to combine multiple similarities of small molecules and miRNAs in future work, which may be more helpful for improving the performance of the model.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jcim.0c00975.

> Baseline method; details of association data and similarity data (Table S1); average recall, specificity, and *G*_mean of four methods within top *k* cutoffs (Table S2); parameter sensitivity analysis (Figure S1); comparison of SMMART with the other methods based on 80%, 70%, 60%, and 50% known associations (Figures S2−S5); performance comparison between SMMART and SMMART*: (a) recall of SMMART and SMMART* within top 5−30% and (b) *G*_mean of SMMART and SMMART* within top 5−30% (Figure S6); average recall, specificity, *G*_mean, and precision of the six methods based on 90%, 80%, 70%, 60%, and 50% known associations within top *k* cutoffs (Table S3−S7); average recall, specificity, *G*_mean, and precision based on association data at different top *k* cutoffs (Table S8); top 10 potential disease candidates discovered by SMMART for the two selected diseases (Table S9) (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Jiawei Luo − *College of Computer Science and Electronic Engineering, Hunan University, Changsha 410083, China*; Email: luojiawei@hnu.edu.cn

Authors

Cong Shen − *College of Computer Science and Electronic Engineering, Hunan University, Changsha 410083, China*; orcid.org/0000-0001-8505-6406

Wenjue Ouyang − *College of Computer Science and Electronic Engineering, Hunan University, Changsha 410083, China*

Pingjian Ding − *School of Computer Science, University of South China, Hengyang 421001, China*; orcid.org/0000-0002-2613-2496

Hao Wu − *College of Computer Science and Electronic Engineering, Hunan University, Changsha 410083, China*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jcim.0c00975

Funding

This work has been supported by the National Natural Science Foundation of China (Grant nos. 61873089, 62032007, and 62002154), the Hunan Provincial Innovation Foundation for Postgraduate (Grant no. CX20200436), and Scientific Research Startup Foundation of University of South China (Grant no. 190XQD096).

Notes

The authors declare no competing financial interest.

■ REFERENCES

(1) Bartel, D. P. MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell* **2004**, *116*, 281−297.

(2) Chen, X.; Xie, D.; Zhao, Q.; You, Z.-H. MicroRNAs and complex diseases: from experimental results to computational models. *Briefings Bioinf.* **2019**, *20*, 515−539.

(3) Ambros, V.; Chen, X. The Company of Biologists Ltd., 2007.

(4) Spizzo, R.; Nicoloso, M. S.; Croce, C. M.; Calin, G. A. SnapShot: microRNAs in cancer. *Cell* **2009**, *137*, 586.

(5) Wang, Y.; Lee, C. G. MicroRNA and cancer−focus on apoptosis. *J. Cell. Mol. Med.* **2009**, *13*, 12−23.

(6) Liu, Z.; Sall, A.; Yang, D. MicroRNA: an emerging therapeutic target and intervention tool. *Int. J. Mol. Sci.* **2008**, *9*, 978−999.

(7) Roshan, R.; Ghosh, T.; Scaria, V.; Pillai, B. MicroRNAs: novel therapeutic targets in neurodegenerative diseases. *Drug Discovery Today* **2009**, *14*, 1123−1129.

(8) Mishra, P. K.; Tyagi, N.; Kumar, M.; Tyagi, S. C. MicroRNAs as a therapeutic target for cardiovascular diseases. *J. Cell. Mol. Med.* **2009**, *13*, 778−789.

(9) Liang, C.; Li, Y.; Luo, J.; Zhang, Z. A novel motif-discovery algorithm to identify co-regulatory motifs in large transcription factor and microRNA co-regulatory networks in human. *Bioinformatics* **2015**, *31*, 2348−2355.

(10) Liu, Y.; Luo, J.; Ding, P. Inferring MicroRNA Targets Based on Restricted Boltzmann Machines. *IEEE J. Biomed. Health Inf.* **2019**, *23*, 427−436.

(11) Pan, C.; Luo, J.; Zhang, J. Computational identification of RNA-Seq based miRNA-mediated prognostic modules in cancer. *IEEE J. Biomed. Health Inf.* **2019**, 626.

(12) Bose, D.; Jayaraj, G.; Suryawanshi, H.; Agarwala, P.; Pore, S. K.; Banerjee, R.; Maiti, S. The tuberculosis drug streptomycin as a potential cancer therapeutic: inhibition of miR-21 function by directly targeting its precursor. *Angew. Chem., Int. Ed.* **2012**, *51*, 1019−1023.

(13) Srinivasan, S.; Selvan, S. T.; Archunan, G.; Gulyas, B.; Padmanabhan, P. MicroRNAs-the next generation therapeutic targets in human diseases. *Theranostics* **2013**, *3*, 930−942.

(14) Hesse, M.; Arenz, C. miRNAs as novel therapeutic targets and diagnostic biomarkers for Parkinson's disease: A patent evaluation of WO2014018650. *Expert Opin. Ther. Pat.* **2014**, *24*, 1271−1276.

(15) Avorn, J. The $2.6 billion pill—methodologic and policy considerations. *N. Engl. J. Med.* **2015**, *372*, 1877−1879.

(16) Sun, M.; Zhao, S.; Gilvary, C.; Elemento, O.; Zhou, J.; Wang, F. Graph convolutional networks for computational drug development and discovery. *Briefings Bioinf.* **2020**, *21*, 919−935.

(17) Ezzat, A.; Wu, M.; Li, X.-L.; Kwoh, C.-K. Computational prediction of drug-target interactions using chemogenomic approaches: an empirical survey. *Briefings Bioinf.* **2019**, *8*, 1337.

(18) Gao, K. Y.; Fokoue, A.; Luo, H.; Iyengar, A.; Dey, S.; Zhang, P. In *Interpretable Drug Target Prediction Using Deep Neural Representation*, Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence; IJCAI, 2018; pp 3371−3377.

(19) Zeng, X.; Zhu, S.; Liu, X.; Zhou, Y.; Nussinov, R.; Cheng, F. deepDR: a network-based deep learning approach to in silico drug repositioning. *Bioinformatics* **2019**, *35*, 5191−5198.

(20) Xuan, P.; Cao, Y.; Zhang, T.; Wang, X.; Pan, S.; Shen, T. Drug repositioning through integration of prior knowledge and projections of drugs and diseases. *Bioinformatics* **2019**, 4108.

(21) Yang, J.; Li, A.; Li, Y.; Guo, X.; Wang, M. A novel approach for drug response prediction in cancer cell lines via network representation learning. *Bioinformatics* **2019**, *35*, 1527−1535.

(22) Chen, X.; Guan, N.-N.; Sun, Y.-Z.; Li, J.-Q.; Qu, J. MicroRNA-small molecule association identification: from experimental results to computational models. *Briefings Bioinf.* **2020**, *21*, 47−61.

(23) Velagapudi, S. P.; Gallo, S. M.; Disney, M. D. Sequence-based design of bioactive small molecules that target precursor microRNAs. *Nat. Chem. Biol.* **2014**, *10*, 291−297.

(24) Disney, M. D.; Winkelsas, A. M.; Velagapudi, S. P.; Southern, M.; Fallahi, M.; Childs-Disney, J. L. Inforna 2.0: A platform for the sequence-based design of small molecules targeting structured RNAs. *ACS Chem. Biol.* **2016**, *11*, 1720−1728.

(25) Xie, W.; Yan, H.; Zhao, X.-M. EmDL: Extracting miRNA-Drug Interactions from Literature. *IEEE/ACM Trans. Comput. Biol. Bioinf.* **2019**, *16*, 1722−1728.

(26) Chen, X.; Xie, W.-B.; Xiao, P.-P.; Zhao, X.-M.; Yan, H. mTD: A database of microRNAs affecting therapeutic effects of drugs. *J. Genet. Genomics* **2017**, *44*, 269−271.

(27) Chen, X.; Sun, Y.-Z.; Zhang, D.-H.; Li, J.-Q.; Yan, G.-Y.; An, J.-Y.; You, Z.-H. J. D. NRDTD: a database for clinically or experimentally supported non-coding RNAs and drug targets associations. *Database* **2017**, *2017*, No. bax057.

(28) Wang, C.-C.; Chen, X.; Qu, J.; Sun, Y.-Z.; Li, J.-Q. modeling, RFSMMA: a new computational model to identify and prioritize potential small molecule-miRNA associations. *J. Chem. Inf. Model.* **2019**, *59*, 1668−1679.

(29) Lv, Y.; Wang, S.; Meng, F.; Yang, L.; Wang, Z.; Wang, J.; Chen, X.; Jiang, W.; Li, Y.; Li, X. Identifying novel associations between small molecules and miRNAs based on integrated molecular networks. *Bioinformatics* **2015**, *31*, 3638−3644.

(30) Li, J.; Lei, K.; Wu, Z.; Li, W.; Liu, G.; Liu, J.; Cheng, F.; Tang, Y. Network-based identification of microRNAs as potential pharmacogenomic biomarkers for anticancer drugs. *Oncotarget* **2016**, *7*, 45584−45596.

(31) Guan, N.-N.; Sun, Y.-Z.; Ming, Z.; Li, J.-Q.; Chen, X. Prediction of potential small molecule-associated microRNAs using graphlet interaction. *Front. Pharmacol.* **2018**, *9*, No. 1152.

(32) Zhao, Y.; Chen, X.; Yin, J.; Qu, J. SNMFSMMA: using symmetric nonnegative matrix factorization and Kronecker regularized least squares to predict potential small molecule-microRNA association. *RNA Biol.* **2020**, *17*, 281−291.

(33) Qu, J.; Chen, X.; Sun, Y.-Z.; Zhao, Y.; Cai, S.-B.; Ming, Z.; You, Z.-H.; Li, J.-Q. In Silico prediction of small molecule-miRNA associations based on the HeteSim algorithm. *Mol. Ther.—Nucleic Acids* **2019**, *14*, 274−286.

(34) Yin, J.; Chen, X.; Wang, C.-C.; Zhao, Y.; Sun, Y.-Z. Prediction of small molecule-microRNA associations by sparse learning and heterogeneous graph inference. *Mol. Pharmaceutics* **2019**, *16*, 3157−3166.

(35) Zhou, X.; Dai, E.; Song, Q.; Ma, X.; Meng, Q.; Jiang, Y.; Jiang, W. In silico drug repositioning based on drug-miRNA associations. *Briefings Bioinf.* **2019**, 498.

(36) Qu, J.; Chen, X.; Sun, Y.-Z.; Li, J.-Q.; Ming, Z. Inferring potential small molecule−miRNA association based on triple layer heterogeneous network. *J. Cheminf.* **2018**, *10*, No. 30.

(37) 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*, 5281−5293.

(38) Pahor, M.; Chrischilles, E.; Guralnik, J.; Brown, S.; Wallace, R.; Carbonin, P. Drug data coding and analysis in epidemiologic studies. *Eur. J. Epidemiol.* **1994**, *10*, 405−411.

(39) Zhou, M.; Chen, Y.; Xu, R. A Drug-Side Effect Context-Sensitive Network approach for drug target prediction. *Bioinformatics* **2019**, *35*, 2100−2107.

(40) Cheng, F.; Kovács, I. A.; Barabási, A.-L. Network-based prediction of drug combinations. *Nat. Commun.* **2019**, *10*, No. 1197.

(41) Ding, P.; Yin, R.; Luo, J.; Kwoh, C. K. Ensemble Prediction of Synergistic Drug Combinations Incorporating Biological, Chemical, Pharmacological and Network Knowledge. *IEEE J. Biomed. Health Inf.* **2018**, *23*, 1336−1345.

(42) Wishart, D. S.; Feunang, Y. D.; Guo, A. C.; Lo, E. J.; Marcu, A.; Grant, J. R.; Sajed, T.; Johnson, D.; Li, C.; Sayeeda, Z.; et al. DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res.* **2018**, *46*, D1074−D1082.

(43) Ding, P.; Shen, C.; Lai, Z.; Liang, C.; Li, G.; Luo, J. Incorporating Multisource Knowledge To Predict Drug Synergy Based on Graph Co-regularization. *J. Chem. Inf. Model.* **2020**, *60*, 37−46.

(44) O'Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R. Open Babel: An open chemical toolbox. *J. Cheminf.* **2011**, *3*, No. 33.

(45) Xiao, Q.; Luo, J.; Liang, C.; Cai, J.; Ding, P. A graph regularized non-negative matrix factorization method for identifying microRNA-disease associations. *Bioinformatics* **2018**, *34*, 239−248.

(46) Wang, J. Z.; Du, Z.; Payattakool, R.; Yu, P. S.; Chen, C.-F. A new method to measure the semantic similarity of GO terms. *Bioinformatics* **2007**, *23*, 1274−1281.

(47) Kozomara, A.; Birgaoanu, M.; Griffiths-Jones, S. miRBase: from microRNA sequences to function. *Nucleic Acids Res.* **2019**, *47*, D155−D162.

(48) Liu, B.; Wang, X.; Lin, L.; Dong, Q.; Wang, X. A discriminative method for protein remote homology detection and fold recognition combining Top-n-grams and latent semantic analysis. *BMC Bioinf.* **2008**, *9*, No. 510.

(49) Wang, D.; Wang, J.; Lu, M.; Song, F.; Cui, Q. Inferring the human microRNA functional similarity and functional network based on microRNA-associated diseases. *Bioinformatics* **2010**, *26*, 1644−1650.

(50) Liu, X.; Wang, S.; Meng, F.; Wang, J.; Zhang, Y.; Dai, E.; Yu, X.; Li, X.; Jiang, W. SM2miR: a database of the experimentally validated small molecules' effects on microRNA expression. *Bioinformatics* **2013**, *29*, 409−411.

(51) Davis, A. P.; Grondin, C. J.; Johnson, R. J.; Sciaky, D.; McMorran, R.; Wiegers, J.; Wiegers, T. C.; Mattingly, C. J. The comparative toxicogenomics database: Update 2019. *Nucleic Acids Res.* **2019**, *47*, D948−D954.

(52) Huang, Z.; Shi, J.; Gao, Y.; Cui, C.; Zhang, S.; Li, J.; Zhou, Y.; Cui, Q. HMDD v3. 0: a database for experimentally supported human microRNA−disease associations. *Nucleic Acids Res.* **2019**, *47*, D1013−D1017.

(53) Xuan, P.; Shen, T.; Wang, X.; Zhang, T.; Zhang, W. Inferring disease-associated microRNAs in heterogeneous networks with node attributes. *IEEE/ACM Trans. Comput. Biol. Bioinf.* **2018**, 1019.

(54) Sarwar, B.; Karypis, G.; Konstan, J.; Riedl, J. *Application of Dimensionality Reduction in Recommender System—A Case Study*; University of Minnesota: Department of Computer Science and Engineering: Minnesota, 2000.

(55) Rubinstein, R. Y.; Marcus, R. Efficiency of multivariate control variates in Monte Carlo simulation. *Oper. Res.* **1985**, *33*, 661−677.

(56) Rockafellar, R. T. Lagrange multipliers and optimality. *SIAM Rev.* **1993**, *35*, 183−238.

(57) Chou, C.-H.; Chang, N.-W.; Shrestha, S.; Hsu, S.-D.; Lin, Y.-L.; Lee, W.-H.; Yang, C.-D.; Hong, H.-C.; Wei, T.-Y.; Tu, S.-J.; et al. miRTarBase 2016: updates to the experimentally validated miRNA-target interactions database. *Nucleic Acids Res.* **2015**, *44*, D239−D247.

(58) Hajian-Tilaki, K. Receiver operating characteristic (ROC) curve analysis for medical diagnostic test evaluation. *Caspian J. Intern. Med.* **2013**, *4*, 627.

(59) Hsu, H.; Lachenbruch, P. A. *Paired t Test*; John Wiley & Sons, Ltd., 2008.

(60) Allison, J. D.; Tanavin, T.; Yang, Y.; Birnbaum, G.; Khalid, U. Various Manifestations of 5-Fluorouracil Cardiotoxicity: A Multicenter Case Series and Review of Literature. *Cardiovasc. Toxicol.* **2020**, 437−442.

(61) Wang, Y.; Xu, W.; Zhang, Q.; Bao, T.; Yang, H.; Huang, W.; Tang, H. J. M. Follow-up of blood glucose distribution characteristics in a health examination population in Chengdu from 2010 to 2016. *Medicine* **2018**, *97*, No. e9763.

(62) Zhang, B.; Pan, X. RDX induces aberrant expression of microRNAs in mouse brain and liver. *Environ. Health Perspect.* **2009**, *117*, 231−240.

(63) Li, B.; Hong, P.; Zheng, C.-C.; Dai, W.; Chen, W.-Y.; Yang, Q.-S.; Han, L.; Tsao, S. W.; Chan, K. T.; Lee, N. P. Y.; et al. Identification of miR-29c and its Target FBXO31 as a Key Regulatory Mechanism in Esophageal Cancer Chemoresistance: Functional Validation and Clinical Significance. *Theranostics* **2019**, *9*, 1599.

(64) He, X.; Li, J.; Guo, W.; Liu, W.; Yu, J.; Song, W.; Dong, L.; Wang, F.; Yu, S.; Zheng, Y.; et al. Targeting the microRNA-21/AP1 axis by 5-fluorouracil and pirarubicin in human hepatocellular carcinoma. *Oncotarget* **2015**, *6*, 2302.