Identifying Potential miRNA-Disease Associations Based on an Improved Manifold Learning Framework

YONG TANG, GAOMING LI, YAZHOU WU, AND DONG YI
Department of Health Statistics, College of Preventive Medicine, Army Medical University, Chongqing 400038, China
Corresponding authors: Yazhou Wu (asiawu@tmmu.edu.cn) and Dong Yi (yd_house@163.com)

This work was supported in part by the National Natural Science Foundation of China under Grant 81473068, Grant 81573254, and Grant 81872716 and in part by the Chongqing Graduate Research and Innovation Project under Grant CYB19183.

ABSTRACT Adequate evidence has shown that miRNA-disease interactions are strongly involved in the pathological processes of complex human diseases. However, it is commonly time-consuming and labor-intensive to utilize laboratory biological experiments to reveal unknown miRNA-disease pairs. Since the previously proposed calculation model has more or fewer deficiencies, we developed the semi-supervised method called Hessian Regularized Non-negative Matrix Factorization Method for miRNA-disease Association prediction (HRNMFMDA). This model introduced Hessian regularization into the NMF framework to preserve the local manifold information and increased an $l_{2,1}$-norm penalty term to ensure the feature selection of the coding matrix and an approximate orthogonal constraint to obtain discriminative information. In the model performance evaluation, HRNMFMDA outperformed eight existing models, achieving AUC values of 0.9074, 0.8618, and 0.9044 +/- 0.0080 in the global leave-one-out cross-validation (LOOCV), local LOOCV and 5-fold cross-validation, respectively. In addition, we applied HRNMFMDA to several high-incidence human carcinomas via three kinds of case studies. Almost all predicted miRNAs were confirmed by external databases derived from experimental literature. Therefore, the conclusion can be drawn that HRNMFMDA is reliable for revealing uncovered miRNA-disease pairs.

INDEX TERMS miRNA-disease association, non-negative matrix factorization, Hessian regularization, discriminability, manifold learning.

I. INTRODUCTION

As a series of short non-coding RNAs of $\sim$22 nucleotides with a single strand [1], microRNAs (miRNAs) are widely found in plants [2], animals [3], and certain viruses [4]. During the post-transcriptional period, they repress gene expression by facilitating the degradation of the target miRNAs [5], thereby playing crucial roles in most cellular processes [6], [7], including cell division [8], differentiation [9], development, metabolism, infection, aging [10], apoptosis [11], and signal transduction [12]. Furthermore, accumulating evidence indicates that abnormal levels of miRNAs are strictly related to numerous human diseases [13]. For instance, miRNA 181a-mediated increased signal of stimulation may trigger the progression of human type 1 diabetes [14].

Moreover, hypercholesterolemia had great connections with the increment of miR-223 levels in atheroprone mice [15]. Moreover, studies on Alzheimer’s disease (AD) have revealed that lower expression levels of miR-15 and miR-107 in the hippocampus of AD brains lead to increased phosphorylation of tau protein, progressively inducing the occurrence of AD [16]. Additionally, experiments have demonstrated that miR-21, miR-494, and miR-1973 have encouraging potential as biomarkers for the diagnosis of classical Hodgkin’s lymphoma [17]. In addition, the reduction in miR-122 can repress the RNA replication of HCV in chimpanzees [18]. Accordingly, uncovering the associations between miRNAs and diseases is needed to understand the occurrence and development of human diseases. However, traditional experimental methods, such as microarray profiling [19], quantitative RT-PCR array [20] and sequencing technologies [21], are time-consuming and expensive. Fortunately, abundant
miRNA-related databases have been continuously established with the accumulated results of innumerable biological experiments. Therefore, it is feasible to build new computational models depending on available biological datasets to predict the underlying disease-related miRNAs.

Over the last few years, investigators have achieved significant progress in clarifying uncovered disease-miRNA associations. Most of the existing predictive models depend on the hypothesis that functionally similar miRNAs are predisposed to diseases with similar phenotypes, and vice versa [22, 23]. Jiang et al. [24] utilized cumulative hypergeometric distribution to construct a scoring system by integrating disease phenotype similarities, miRNA functional similarities, and known disease-miRNA associations. However, the prediction performance was not satisfactory since the information of other nodes, except for the miRNAs’ neighbors, was not effectively exploited. Then, Jiang et al. [25] developed a genomic data integration-based naive Bayesian model to predict novel miRNAs, improving predictive outcomes and providing guidance for further experimental investigation. Moreover, Chen et al. [26] proposed Random Walk with Restart (RWR) based on the miRNA functional similarities network (MFSN) to rank candidate miRNAs of a specific disease. However, this method did not work for diseases without any verified miRNAs. The Human Disease-related MiRNA Prediction (HDMP) method was built by Xuan et al. [27] by combining the semantic similarities and phenotype similarities of diseases. Then, miRNA clusters were allocated with higher importance according to the k nearest nodes. Nonetheless, HDMP ignored topology information embedded by local neighbors and could not infer novel miRNAs of the isolated diseases. Later, Xuan et al. [28] further proposed the MIDP method based on Random Walk. In this model, the prior information of topologies was integrated effectively by assigning different weights to known and unknown nodes. Furthermore, the extended transition on the miRNA-disease bilayer network made it possible to implement the prediction of isolated disease-related miRNAs. Chen et al. [29] presented the Heterogeneous Graph Inference (HGI-MDA) method to identify candidate miRNA-disease pairs. The model constructed a target equation based on a heterogeneous graph. Then, matrix multiplications were calculated iteratively until convergence to acquire the predictive probability matrix. Nonetheless, the model may bias those miRNAs with more relevant disease records. Chen et al. [30] developed Matrix Decomposition and Heterogeneous Graph Inference (MDHGI) for the identification of miRNA-disease associations. In this method, the heterogeneous network was built using a fine-tuned associations matrix adjusted by low-rank matrix decomposition. Moreover, MDHGI could achieve improved performance because the model combined matrix decomposition with heterogeneous graph inference. Inevitably, the prediction solution had several defects, such as the difficult problem of parameter optimization and the bias caused by unevenly distributed miRNA-disease pairs. As a network-based model, the Path-Based (PBMDA) predictive model was introduced by You et al. [31] by combining multiple biological datasets. The method used the depth-first search algorithm to determine the weights of each path and prioritized miRNA-disease pairs based on their aggregated scores. Chen et al. [32] presented the Bipartite Network Projection (BNPMDA) for identifying miRNA-disease associations. The method adopted hierarchical clustering to generate bias ratings between miRNAs and diseases. Then, a bipartite network recommendation algorithm was implemented in the weight reassignment step between the investigated disease and other miRNAs. Inevitably, BNPMDA also had a crucial flaw in that it was not suitable for the association prediction of isolated diseases, which limited its range of application. The Ensemble Learning and Link Prediction (ELLPMDA) model [33] combined three classic link prediction methods (common neighbors, Jaccard index and Katz index) based on ensemble learning for identifying miRNA-disease associations. The prediction accuracy was effectively improved. Furthermore, Chen et al. [34] proposed a prediction model based on Bipartite Local models and Hubness-Aware Regression (BLHARMDA). The model used Jaccard similarity and KNN methods to mitigate the adverse effects of ‘bad’ hubs.

Apart from the models mentioned above, many computational methods that belong to machine learning are being gradually created. Xu et al. [35] developed a novel method by utilizing the miRNA target-dysregulated network (MTDN) to discover candidate prostate cancer miRNAs. Adopting the support vector machine (SVM) classifier, Xu’s model made predictions by topological characteristics and expression folds of miRNA. Although this method was creative, it could not achieve outstanding accuracy because of the high false-positive rates of miRNA-target pairs. Chen et al. [36] presented a global learning model based on Regularized Least Squares (RLSMDA) for the candidates’ prediction. Furthermore, it was mainly constructed as a continuous classifier composed of regularized least squares from the perspective of miRNA and disease. Furthermore, RLSMDA could work for solitary diseases or miRNAs. A novel model based on the restricted Boltzmann machine (RBMMMDA) was developed by Chen et al. [37]. It is noteworthy that RBMMMDA was the earliest method that could discover not only the unobserved miRNA-disease pairs but also their association types. Of course, the model had some limitations, e.g., RBMMMDA did not have the applicability to reveal the isolated diseases or miRNAs. Moreover, Liang et al. [38] constructed the $\ell_1$-norm graph model. In this semi-supervised model, an effective $\ell_1$-norm was introduced into the objective function, and then the weights of the miRNA and disease similarity matrix were updated adaptively in the iterative calculation process until convergence. The Extreme Gradient Boosting Machine-based method (EGBMMDA) was introduced by Chen et al. [39]. In this method, informative feature vectors were acquired by combining graph theory features, statistical features, and matrix factorization from multiple biological datasets. Nevertheless, choosing negative samples inappropriately would impair the predictive
performance of the model, which deserves further investigation. Chen et al. [40] built the associations inference based on Inductive Matrix Completion (IMCMDA), whose core idea is to recover the missing entries using an iterative equation from verified miRNA-disease pairs. Chen et al. [41] developed the Ensemble of Decision Tree method (EDT-MDA) for revealing potential candidate pairs. The model adopted principal components analysis (PCA) to reduce feature dimensionality and then utilized the decision trees as base learners to acquire the predictive score matrix. However, the method might cause bias to diseases associated with more miRNAs. Chen et al. [42] developed an improved prediction model based on Laplacian Regularized Sparse Subspace Learning (LRSSLMDA). In this model, statistical and graph-theoretical features of miRNAs and diseases extracted from the integrated similarities were integrated into a common subspace by the L1-norm constraint and Laplacian regularization terms to acquire the projection matrices from the miRNA and disease perspective, respectively. The final prediction scores were obtained by taking the average of predicted association scores of miRNA and disease. Recently, Zhao et al. [43] proposed a novel method based on an Adaptive Boosting algorithm (ABMDA). This model adaptively adjusted the weights of weak classifiers (decision trees) to combine them into a strong classifier. Moreover, ABMDA was suitable for predicting isolated diseases without any known related miRNAs. Wang et al. [44] developed a new prediction method called LMTRDA. It used the Logical Model Tree (LMT) as the base classifier and then adjusted the weights of the corresponding samples. It is worth mentioning that this model uses natural language processing technology to integrate miRNA sequences into multiple data sources, improving the final prediction performance. The RFMDA model, proposed by Chen et al. [45], performed feature extractions of the integrated multiple data sources according to the miRNA-disease pair and then input them into a random forest for training. The model had the disadvantage that it was often difficult to obtain the negative samples that were required during the training process. Wang et al. [46] constructed the Negative Samples Extraction method (NSEMDA) to predict miRNA-disease association pairs. The model used the Spy and Rocchio techniques to exclude negative samples and calculated similarity for suspected negative samples. Finally, the support vector machine-similarity weight method was used to obtain predictive scores. The advantage of the model was that it could mitigate the effects of noisy data. Zhang et al. [47] created a predictive model based on variational autoencoders (VAEMDA). This model used spliced matrices generated from adjacency and similarity matrices as training data and took the average of the outputs of the two VAE models as the final prediction scores. However, this type of model lacked interpretability, which needs to be further improved.

Some progress has been made in previous research on miRNA-disease association predictive models. However, there is still room to further improve the predictive performance of the models. Therefore, we presented an improved solution called Hessian Regularized Non-negative Matrix Factorization Method for miRNA-disease Association Prediction (HRNMFMDA) to further utilize information from negative samples. This model first introduced Hessian regularization into the NMF framework to exploit the local manifold structure by recovering the inherent local information from the training data. Moreover, the method increased the $l_{2,1}$-norm constraint and approximate orthogonal constraints to ensure the feature selection of the coding matrix, thereby improving the final predictive performance. HRNMFMDA acquired AUCs of 0.9074, 0.8618, and 0.9044+/−0.0080 after the implementation of global LOOCV, local LOOCV and 5-fold cross-validation, respectively, which were superior to most of the previously proposed models. Moreover, HRNMFMDA was further applied to infer several common human tumors. With the top 50 miRNAs in the predictive results, most of them were confirmed through dbDEMC [48], miR2Disease [49], and HMDD V3.0 [50]. Therefore, the results of the various verification experiments indicate that HRNMFMDA is useful for identifying candidate miRNA-disease pairs.

II. MATERIALS AND METHODS

A. ASSOCIATIONS DATASET

Over the past decades, cumulative biological experiments have discovered a considerable number of miRNA-disease associated pairs. Here, the known human associations used in this model were downloaded from HMDD V2.0 [51]. The dataset contained 495 miRNAs and 383 diseases. An adjacency matrix was adopted to formalize the 5430 experimentally verified human miRNA-disease associated pairs. Specifically, if the disease $d(i)$ had a previously verified connection with miRNA $m(j)$, $A_{ij}$ was assigned a value of 1; apart from this, the corresponding position was set to 0.

B. MIRNA FUNCTIONAL SIMILARITIES

The miRNA functional similarities of this paper were gathered from http://www.cuilab.cn [23], whose calculations were inspired by a fundamental premise that functional analogous miRNAs more likely had a link with analogous phenotypic diseases [52]. Next, an miRNA functional similarity matrix $FS$ was constructed to store data for this association matrix, where $FS(i, j)$ denotes the functional similarities between the $i$-th and $j$-th miRNA.

C. DISEASE SEMANTIC SIMILARITIES MEASUREMENT 1

As reported by Wang et al. [23], a directed acyclic graph (DAG) was utilized to outline the semantic correlations among diseases. For instance, disease $D$ could be described using a triple structure such as $DAG(D) = (D, T(D), E(D))$. In this expression, the node-set $T(D)$ represents the corresponding disease and all its ancestor nodes, and edge-set $E(D)$ was defined as the directed edges from the parents to the child nodes. From the above, the semantic value of disease $D$
can be defined by:

\[ DV1(D) = \sum_{d \in T(D)} D1_D(d) \]  

(1)

and

\[
\begin{cases}
D1_D(d) = 1, & \text{if } d = D \\
D1_D(d) = \max\{\Delta^s D1_D(d') \mid d' \in \text{children of } d\}, & \text{if } d \neq D
\end{cases}
\]  

(2)

where \( \Delta \) is the semantic decay factor and was set to 0.5 in the iterative equation according to reference [23]. Commonly, the farther a child node is located from the root node, the smaller its contribution to the semantic value of the corresponding disease. Referencing the hypothesis that two diseases sharing a greater degree of overlap of DAGs should have higher semantic similarity, the semantic similarities between the \( i \)-th and \( j \)-th disease could be defined by:

\[ SS1(d(i), d(j)) = \frac{\sum_{t \in T(d(i)) \cap T(d(j))} (D1_{d(i)}(t) + D1_{d(j)}(t))}{DV1(d(i)) + DV1(d(j))} \]

(3)

D. DISEASE SEMANTIC SIMILARITIES MEASUREMENT 2

Considering that the numbers of DAGs involving different disease terms can vary greatly, another method was proposed for quantifying the semantic similarities of the diseases more objectively and accurately [27]. According to the assumption that disease terms appearing in fewer DAGs should be given a more prominent contribution, we have:

\[ D2_D(d) = -\log\left(\frac{\text{the number of DAGs including } d}{\text{the number of disease}}\right) \]

(4)

Similarly, the semantic similarity of disease \( D \) and the semantic similarity between the \( i \)-th and \( j \)-th diseases can be defined as follows:

\[ DV2(D) = \sum_{d \in T(D)} D2_D(d) \]

(5)

and

\[ SS2(d(i), d(j)) = \frac{\sum_{t \in T(d(i)) \cap T(d(j))} (D2_{d(i)}(t) + D2_{d(j)}(t))}{DV2(d(i)) + DV2(d(j))} \]

(6)

E. GAUSSIAN SIMILARITIES FOR DISEASES AND MiRNAs

According to the literature [53], the radial basis function was adopted to measure the Gaussian similarities between disease \( d(i) \) and disease \( d(j) \). Specifically, we used the binary interaction profile vectors \( IP(d(i)) \) and \( IP(d(j)) \) to represent the \( i \)-th row and the \( j \)-th row of the miRNA-disease adjacency matrix, respectively. Then the disease Gaussian similarities matrix \( KD \) was calculated by:

\[ KD(d(i), d(j)) = \exp(-\gamma_d \|IP(d(i)) - IP(d(j))\|^2) \]

(7)

where \( \gamma_d \) is the bandwidth parameter, which is calculated as follows:

\[ \gamma_d = \frac{1}{md} \sum_{i=1}^{md} \|IP(d(i))\|^2 \]

(8)

Analogously, Gaussian similarities of miRNAs were determined as follows:

\[ KM(m(i), m(j)) = \exp(-\gamma_m \|IP(m(i)) - IP(m(j))\|^2) \]

(9)

and

\[ \gamma_m = \frac{1}{nm} \sum_{i=1}^{nm} \|IP(m(i))\|^2 \]

(10)

For brevity, the values of \( \gamma_d \) and \( \gamma_m \) were both assigned a value of 1, consistent with the previous literature [53], [54].

F. INTEGRATED SIMILARITIES FOR DISEASES AND MiRNAs

Typically, disease semantic similarities were available if the corresponding DAGs existed. If not, the same method would not work. To acquire the disease similarities without missing data, the semantic similarities of the diseases were combined with the Gaussian interaction profile kernel similarities by piecewise function referring to the literature [55]:

\[ S_d(d(i), d(j)) = \begin{cases} \frac{SS1(d(i), d(j)) + SS2(d(i), d(j))}{2} & \text{if } d(i) \text{ and } d(j) \text{ have semantic similarity} \\ KD(d(i), d(j)), & \text{otherwise} \end{cases} \]

(11)

where taking the arithmetic mean as the disease semantic similarity if the DAGs existed was adopted; otherwise, KD. Similarly, the integration of miRNA similarity was obtained as follows:

\[ S_m(m(i), m(j)) = \begin{cases} \frac{FS(m(i), m(j))}{2} & \text{if } m(i) \text{ and } m(j) \text{ have semantic similarity} \\ KM(m(i), m(j)), & \text{otherwise} \end{cases} \]

(12)

G. HESSIAN REGULARIZED NON-NEGATIVE MATRIX FACTORIZATION WITH \( \ell_2,1 \)-NORM AND DISCRIMINABILITY FOR DISEASE-ASSOCIATED MiRNAs PREDICTION

1) STANDARD NMF

NMF is a promising and effective decomposition method utilized in various situations, including document analysis [56], pattern recognition [57], and image processing [58]. In the NMF, given the miRNA-disease association matrix \( A \in \mathbb{R}^{m \times n} \), the target of the technique is to seek \( P \in \mathbb{R}^{m \times r} \) and \( Q \in \mathbb{R}^{r \times n} \) by minimizing the following:

\[
\min_{P,Q} \frac{1}{2} \| A - PQ \|_F^2, \quad P_{ik} \geq 0, \quad Q_{kj} \geq 0, \quad \forall i, j, k
\]

(13)
where \(\| \cdot \|_F\) represents the Frobenius norm. The minimization of the equation above could be solved by an iterative updating algorithm proposed by Lee and Seung [59].

2) HESSIAN REGULARIZATION

Inspired by the literature [60], the expression of Hessian energy \(S_H(f)\) could be defined by the integral of the volume element \(dV(x)\) of a smooth manifold \(M \subset \mathbb{R}^n\), as follows:

\[
S_H(f) = \int_M \| \nabla_a \nabla_b f \|_{T_x M}^2 dV(x) \tag{14}
\]

where \(T_x M\) represented point \(x\) of the tangent space of \(M\), and \(\nabla_a \nabla_b f\) denoted the second covariant derivation of function \(f\), which was a real-valued mapping: \(M \rightarrow \mathbb{R}^m\) [60]. Using the conventional coordinate system, we further had:

\[
\int_M \| \nabla_a \nabla_b f \|_{T_x M}^2 = \sum_{r,s=1}^d \left( \frac{\partial^2 f}{\partial C_r \partial C_s} \right)^2 \tag{15}
\]

where \(C_r\) and \(C_s\) were normal coordinates. According to this, the second-order covariant derivative of function \(f\) happened to be the Frobenius norm expression of the Hessian regularization in standard coordinates. Thus, we obtained:

\[
S_{Hess}(f) = \sum_{i=1}^n \sum_{r,s=1}^d \left( \frac{\partial^2 f}{\partial C_r \partial C_s} |_{x_i} \right)^2 \tag{16}
\]

Using the least squares method to solve a second-order polynomial could obtain the \(H\) operator [61]. Therefore, the approximate representation of the Frobenius norm of Hessian regularized \(f\) on \(x_i\) could be expressed as:

\[
\| \nabla_a \nabla_b f \|_2^2 \approx \sum_{r,s=1}^d \left( \sum_{\alpha=1}^k H_{ij}^{(i)}(\alpha)^2 \right) = \sum_{r,s=1}^d \left( \sum_{\alpha=1}^k f(\alpha) f(\beta) B_{ij}^{(i)} \right) \tag{17}
\]

where \(B_{ij}^{(i)} = \sum_{r,s=1}^d H_{ij}^{(i)} H_{rs}^{(j)}\) and the final estimated Hessian expression \(\hat{S}_{Hess}(f)\) could be approximated as follows:

\[
\hat{S}_{Hess}(f) = \sum_{i=1}^n \sum_{r,s=1}^d \left( \frac{\partial^2 f}{\partial C_r \partial C_s} |_{x_i} \right)^2 \tag{18}
\]

\[
= \sum_{i=1}^n \sum_{\alpha \in N_k(x_i)} \sum_{\beta \in N_k(x_i)} f_{\alpha \beta} B_{ij}^{(i)} \tag{19}
\]

\[
= \langle f, B f \rangle
\]

where \(n\) is the number of samples, and \(B\) denotes the Hessian regularization, which is made up of the sum of all matrices \(B^{(i)}\) [62], [63]. The specific calculation process of Hessian regularization is shown in Algorithm 1.

Algorithm 1 Solving the Hessian Regularization Term

Input:

- Similarity matrix \(S\), Number of neighbors \(k\), \(d\) columns of matrix \(U\) after the PCA decomposition

Procedure:

1. For each sample element \(x_i\) in \(S\), find its \(k\) nearest neighbors \(N_k\) then form a matrix \(X_i\). For all \(x_j \in N_k\), each row is composed of the centralized \(x_j - x_i\);
2. Perform singular value decomposition \(X_i = UDV^T\) on \(X_i\) to estimate the orthogonal coordinate of the tangent space;
3. Perform Gram-Schmidt orthogonal normalization on matrix \(M = [1 \ U_1 \ldots U_d \ U_{11} \ldots U_{dd}]\) and denote the result as \(H\). The Frobenius norm of \(H\) is \(H^T H\);
4. Calculate the sum of \(H^T H\) in all instances, and finally obtain the Hessian regularization term \(B\).

Output: Hessian regularization term \(B\)

3) HRNMFDMA

Motivated by previous literature [64], [65], sparseness constraints [66] and discriminative constraints [67] were joined into Hessian regularized NMF. In general, the \(\ell_{2,1}\)-norm term of the matrix \(Q\) and orthogonal constraint terms on matrix \(Q\) were added to the NMF framework as penalty terms. Then, the objective equation of HRNMFDMA to be optimized could be given as follows:

\[
\min_{P,Q} \frac{1}{2} \| A - PQ \|_F^2 + \lambda_1 tr(QB_mQ^T) + \lambda_2 tr(PB_dP^T)
\]

\[
+ \mu \| Q^T \mathbf{I} - Q \|_{2,1} + \gamma_1 \| Q \|_{2,1} + \gamma_2 \| P \|_2^2.
\]

s.t. \(P_{ik} \geq 0\), \(Q_{kj} \geq 0\), \(\forall i, j, k\) \(\tag{20}\)

In the equation above, \(\| A - PQ \|_F^2\) is the reconstruction error term of the non-negative matrix factorization, also known as the least squares cost function, where matrix \(A\) represents the miRNA-disease interactions and matrices \(P\) and \(Q\) are the basic matrix and encoding matrix to be solved, respectively.

\(tr(PB_dP^T)\) and \(tr(QB_mQ^T)\) are the Hessian regularization terms for preserving the manifold information, where \(tr(M)\) denotes the trace of the matrix \(M\).

\[
\| Q \|_{2,1} = \sum_{j=1}^k |q_{ij}|\]

where \(q_{ij}\) represented the \(j\)-th row of matrix \(Q\). The purpose of row sparse regularization was to retain pivotal features only by decreasing some of the rows of matrix \(Q\) to 0.

\[
\| Q^T \mathbf{I} - Q \|_{2,1} = \| Q^T \mathbf{I} - Q \|_1
\]

was the discriminant constraint term, where \(\mathbf{I}\) was a unit matrix with the dimensions \(k \times k\), whose purpose was to make matrix \(Q\) approximately orthogonal to obtain discriminant information.

The Tikhonov (L2) regularization term \(\| P \|_2^2\) was set to ensure the smoothness of matrix \(P\) to prevent overfitting.
Among them, \( \lambda, \mu, \gamma \) were regularization parameters, which could be further optimized using separate validation sets. For the sake of brevity, default parameters were used in the following sections.

4) OPTIMIZATION

Since Eq. (20) was non-convex, the closed-form solution was not available. Next, an iterative algorithm was utilized to obtain the local minimum instead. According to formula \( \|A\|_F = tr(AA^T) \), the optimization function can be rewritten in the following form:

\[
O = \frac{1}{2} tr([AA^T - 2AQ^TP + PQQ^TP^T] + \lambda_1 tr(QB_mQ^T) + \lambda_2 tr(PB_dP^T) + \mu \|Q^2 - 1\|_F + \gamma_\|Q\|_{2,1} + \gamma_2 \|P\|_2^2, \quad s.t. P_{ik} \geq 0, \quad Q_{kj} \geq 0, \forall i, j, k
\]

The corresponding Lagrange function \( L \) of the above equation can be defined as:

\[
L = \frac{1}{2} tr([AA^T - 2AQ^TP + PQQ^TP^T] + \lambda_1 tr(QB_mQ^T) + \lambda_2 tr(PB_dP^T) + \mu \|Q^2 - 1\|_F + \gamma_\|Q\|_{2,1} + \gamma_2 \|P\|_2^2, \quad s.t. P_{ik} \geq 0, \quad Q_{kj} \geq 0, \forall i, j, k
\]

where \( \psi_{ik} \) and \( \Phi_{ik} \) denote the Lagrange multipliers for the constraints \( P_{ik} \geq 0 \) and \( Q_{kj} \geq 0 \), respectively. The partial derivatives of the equation above concerning the matrix \( P \) were:

\[
\frac{\partial L}{\partial P} = PQQ^T - AQ^T + 2\lambda_2 B_d P + \gamma_2 P + \psi
\]

Following the Karush-Kuhn-Tucker (KKT) conditions [68] and \( \psi_{ik} P_{ik} = 0, \) so

\[
(PQQ^T - AQ^T + 2\lambda_2 B_d P - 2\lambda_2 B_d P + \gamma_2 P)_{ik} P_{ik} = 0
\]

where \( B_d = B_d^+ - B_d^-, \quad B_d^+ = \frac{B_d^+ + B_d^-}{2}, \quad B_d^- = \frac{B_d^+ - B_d^-}{2}
\]

From the above, the updating rules were acquired as follows:

\[
P_{ik} = \frac{(AQ^T + 2\lambda_2 B_d P)}{(PQQ^T + 2\lambda_2 B_d P + \gamma_2 P)_{ik}}
\]

Similarly, taking the partial derivatives of \( L \) above concerning matrix \( Q \), we could obtain:

\[
\frac{\partial L}{\partial Q} = PQ^T - P^T A + 2\lambda_1 Q B_m + 4\mu QQ^T Q - 4\mu Q + \gamma_1 R Q + \Phi
\]

Here, \( R \) was a diagonal matrix whose \( i \)-th element was:

\[
R_{ii} = \frac{1}{2 \|Q_{i}\|_2^2}
\]

Then, using KKT conditions, \( \Phi_{ik} Q_{kj} = 0 \), we can obtain:

\[
(P^T Q - P^T A + 2\lambda_1 Q B_m^+ - 2\lambda_1 Q B_m^- + 4\mu QQ^T Q - 4\mu Q + \gamma_1 R Q)_{kj} = 0
\]

where \( B_m = B_m^+ - B_m^- \), \( B_m^+ = \frac{B_m^+ + B_m^-}{2} \), \( B_m^- = \frac{B_m^+ - B_m^-}{2} \)

Finally, the iterative rule for matrix \( Q \) was obtained from the above as follows:

\[
Q_{kj} = \frac{(P^T A + 2\lambda_1 Q B_m^+ + 4\mu QQ^T Q + \gamma_1 R Q)_{kj}}{(P^T Q + 2\lambda_1 Q B_m^+ + 4\mu QQ^T Q + \gamma_1 R Q)_{kj}}
\]

The main goal of HRNMFMDA was to recover the complete miRNA-disease association matrix \( X_{pred} \in \mathbb{R}^{m \times n} \) by implementing NMF with Hessian regularization on the known entries of miRNA-disease associations matrix \( X \in \mathbb{R}^{m \times n} \). In other words, the main problem we needed to solve was to find \( P \in \mathbb{R}^{m \times r} \) and \( Q \in \mathbb{R}^{r \times n} \) to acquire the formula \( X_{pred} = PQ \), where the parameter \( r \) represented the theoretical rank constrained with \( \min(\text{rank}(P), \text{rank}(Q)) \). The value of \( r \) would influence the convergence speed of the NMF factorization, but it had minimal effect on the final results [40], [69]. Expressly, the value of \( r \) was set as 100, the convergence threshold was set as \( 10^{-6} \), and the above regularization parameter values take \( \lambda_1 = \lambda_2 = 1, \mu = 1, \gamma_1 = \gamma_2 = 1 \) for simplicity. After preliminary exploration, the parameter settings of the Hessian regularization were \( k = 30 \) and \( d = 10 \) in Algorithm 1. After updating \( P \) and \( Q \) based on Eq. (25) and Eq. (29) until convergence, we obtained the final prediction score matrix as \( X_{pred} = PQ \). Finally, we ranked the corresponding miRNAs based on \( X_{pred} \). Generally, the miRNAs at the top of the prediction scores had a higher probability of being related to the corresponding disease. The flow diagram of HRNMFMDA is depicted in Figure 1 and the overall implementation of the model is shown in Algorithm 2. The dataset and codes of this model can be downloaded through the following GitHub address: https://github.com/Tomhappy/HRNMFMDA.

III. RESULTS

A. PERFORMANCE VALIDATIONS

In this section, LOOCV validation and five-fold cross-validation were performed based on HMDD V2.0 to assess the predictive performance of HRNMFMDA. LOOCV validation comprised two subtypes, including global and local LOOCV. Global LOOCV focused on potential miRNA-disease associations for all diseases. In contrast, local LOOCV only considered miRNAs that target specific diseases. During the implementation, we removed the known miRNA-disease pair one by one as a test sample (total 5430 verification tests) one by one, regarding the other known pairs as training samples and all remaining unverified associations as candidates. Specifically, global LOOCV concentrated on all of the undiscovered associations, while local LOOCV only focused on these miRNAs related to the specific investigated disease. For these two kinds of LOOCV, test samples were ranked with candidates according to HRNMFMDA, and exceeding a predetermined threshold indicated a successful prediction and vice versa. Then, with the changing threshold, the receiver operating characteristic (ROC) curve could
be plotted. And the area under the ROC curve (AUC) was employed to evaluate the prediction performance of the classifier. AUC = 1 indicated the best theoretical performance was achieved, while 0.5 indicated random performance.

To check the efficacy of HRNMFMDA, the model was compared with a series of competitive methods, whose parameter settings were all chosen according to the recommendation in the original literature, specifically: $k = 50$, $\eta_1 = 0.1$, $p = 5$, $\lambda_m = \lambda_d = 0.01$ for GRNMF [70]; decay factor $\alpha = 0.5$ for HGIMDA [29]; $\lambda_1 = \lambda_2 = 1$, convergence criterion $= 10^{-6}$ for IMCMDA [40]; $\eta_M = \eta_D = 1$, $w = 0.9$ for RLSMDA [36]; $q = 170$, $p = 220$, $r = 180$ for GRMDA [71]; k-nearest-neighbors $k = 2$ for RKNMMDA [72]; and $r = 0.2$, cutoff $= 10^{-6}$ for RWRMDA [26]. Moreover, all the above methods were based on the same dataset HMDD V2.0. As shown in Figure 2, HRNMFMDA achieved AUCs of 0.9074 and 0.8618 in global and local LOOCV, respectively. Correspondingly, GRNMF, HGIMDA, IMCMDA, RLSMDA, GRMDA, WBSMDA [55] and RKNMMDA achieved AUCs of 0.8869, 0.8622, 0.8379, 0.8377, 0.8272, 0.7998 and 0.7222 in global LOOCV, respectively. In local LOOCV, these models obtained AUCs of 0.8577, 0.8322, 0.8276, 0.5787, 0.8160, 0.8013 and 0.7307, respectively. These results confirmed that HRNMFMDA achieved superior performance for identifying uncovered miRNA-disease associations. In addition, RWRMDA scored an AUC of 0.7808 in local LOOCV. RWRMDA was not compared in global validation due to its inability to infer potential miRNAs for all diseases simultaneously. More importantly, RLSMDA yielded a low AUC value in local LOOCV, which was likely due to the apparent sparsity of the dataset used in our experiments that compromised its predictive performance.

The verified pairs, as mentioned above, were partitioned into five equal-sized parts randomly in the 5-fold cross-validation part. Each subgroup was then treated as the test set in turn, and the remaining subsets were considered training cases. It is worth mentioning that each cross-validation used the updated Gaussian kernel similarities based on the training data. Next, using the score matrix generated by HRNMFMDA, we ranked the miRNAs of the test set compared with the candidates. This calculation process was repeated 100 times to accomplish the stable estimates of the mean and variance of HRNMFMDA. After the implementation of validation, we acquired the average AUCs of GRNMF, HGIMDA, IMCMDA, RLSMDA, GRMDA, WBSMDA and RKNMMDA as 0.8689+/−0.00065, 0.8594+/−0.0021, 0.8335+/−0.00056, 0.8351+/−0.00048, 0.8056+/−0.0022, 0.8009+/−0.0012 and 0.6761+/−0.0020, respectively. Our proposed method, HRNMFMDA achieved an average AUC of 0.9044+/−0.0080. The results further proved the reliability of HRNMFMDA. A local ranking method, RWRMDA was not included in this global validation framework. All these results indicated that HRNMFMDA achieved high prediction accuracy and good robustness.

As described above, the ranking results for 5430 test samples were obtained during the global and local LOOCV. Then, the paired t-test was implemented to assess whether there was a significant difference in the performance of HRNMFMDA between several other methods (GRNMF, HGIMDA, IMCMDA, RLSMDA, GRMDA, WBSMDA, RKNMMDA and RWRMDA) using the p-value. Finally, we concluded that HRNMFMDA differed significantly from other algorithms (see Table 1).

## B. STRUCTURE ANALYSIS

To evaluate the performance contribution of the Hessian regularization term to the model in our study, we replaced the Hessian regularization in HRNMFMDA with Laplacian regularization and kept the other parts of the model unchanged. The AUC values corresponding to the two methods performed in 5 types of cross-validations are shown in Table 2. These results suggested that prediction using either of these two kinds of regularization could produce satisfactory performance, but Hessian regularization yielded better predictive performance than Laplacian regularization. The above results agreed with the existing theoretical research conclusions that

### Algorithm 2 Overall Implementation of HRNMFMDA

**Input:**
- DAGs of diseases, denoted as $DAG(D) = (D, T(D), E(D))$, miRNA functional similarity matrix $FS$, known miRNA-disease associated pairs matrix $A$, parameters and default values
- $\gamma_d = \gamma_m = 1$, $\lambda_1 = \lambda_2 = 1$, $\mu = 1$, $\gamma_1 = \gamma_2 = 1$, $k = 30$, $d = 10$

**Procedure:**
1. For each disease $d(i)$ in the association matrix $A$
   - If there are corresponding DAGs
     - Calculate the two semantic similarities $SS1$ and $SS2$ according to Eq. 1 to 6.
   - Else
     - Calculate the Gaussian kernel similarity $KD$ according to Eq. 7 and 8 (about parameter $\gamma_d$).
2. For each miRNA $m(j)$ in the association matrix $A$
   - Calculate the Gaussian kernel similarity $KM$ and $SS1/m$, respectively, and calculate Hessian regularization matrices $B_d$ and $B_m$ (about parameters $k$ and $d$).
3. According to Algorithm 1, input similarity matrices $S_d$ and $S_m$ respectively, and calculate Hessian regularizations $B_d$ and $B_m$.
4. According to Eq. 25 and Eq. 29, iteratively update the matrix to obtain $P$ and $Q$ (about the parameters $\gamma_1$, $\gamma_2$, $\mu$, $\gamma_1$ and $\gamma_2$) until the maximum number of iterations or convergence;
5. Calculate $X_{pred} = PQ$

**Output:** Predicted score matrix $X_{pred}$
the second-order Hessian energy could overcome the problem that the graph Laplacian lacks extrapolating ability [64], [73]. In summary, Hessian regularization had an advantage in discovering unverified miRNA-disease associations compared to Laplacian regularization.

Finally, the permutation test was performed to evaluate how much the two integrated similarities matrices and adjacency matrix contributed to the predictive performance. The core approach of the test randomized one of the three matrices and kept the other two unchanged in the 5-fold cross-validation framework (LOOCV validation framework was not adopted here for the calculations would take too long, even several months). The test for each input matrix was randomized and repeated 100 times, and the average
Identifying Potential miRNA-Disease Associations

**FIGURE 2.** Performance comparison was implemented between HRNMFMDA and eight previous methods (GRNMF, HGIMDA, IMCMDA, RLSMDA, GRMDA, WBSMDA, RKNNMDA, and RWRMDA) in terms of ROC curves and AUCs. As a result, HRNMFMDA achieved an AUC of 0.9074 in global LOOCV and an AUC of 0.8618 in local LOOCV, outperforming other previous models.

**TABLE 1.** A comparison of the paired t-test was implemented to validate the performance difference between HRNMFMDA and other previous models based on the predicted positions.

| Comparison methods | p-value based on global LOOCV | p-value based on local LOOCV |
|--------------------|-------------------------------|-------------------------------|
| GRNMF              | 8.5200e-28                   | 2.405333e-06                 |
| HGIMDA             | 3.9614e-73                   | 1.529746e-31                 |
| IMCMDA             | 1.0236e-128                  | 1.240057e-44                 |
| RLSMDA             | < 2.2e-16                    | < 2.2e-16                    |
| GRMDA              | 9.3594e-118                  | 6.290587e-47                 |
| WBSMDA             | < 2.2e-16                    | 7.320137e-16                 |
| RKNNMDA            | < 2.2e-16                    | 2.964394e-322                |
| RWRMDA             | —                             | < 2.2e-16                    |

**TABLE 2.** To evaluate the predictive contribution of the regularization terms in HRNMFMDA, Hessian regularization and Laplacian regularization were utilized separately for prediction in the following three cross-validations. The detailed AUCs are listed in the following table.

| Manifold Regularization Term | Global LOOCV | Local LOOCV | 5-fold CV |
|------------------------------|--------------|-------------|-----------|
| Hessian                      | 0.9074       | 0.8618      | 0.9044+/-.00080 |
| Laplacian                    | 0.9043       | 0.8562      | 0.9011+/-.00033 |

AUCs were taken as the test result. If a particular matrix made a more significant contribution to the predictive ability of the method, the permutation test results based on that matrix would be closer to 0.5. As described in Table 3, HRNMFMDA randomizing of the adjacency matrix had the average AUCs much closer to 0.5 than that of the others, indicating that the miRNA-disease association was of more paramount importance. In addition, the disease similarities made more significant contributions than the miRNA similarities for improving predictive power.

**C. CASE STUDIES**

Next, three kinds of case studies were implemented to further verify the predictive capability of HRNMFMDA for uncovering novel miRNA-disease pairs. The first type of case study included breast neoplasms, utilizing the verified pairs of HMDD V2.0 [51] as the training samples. For every predicted disease, the corresponding unverified miRNAs were ranked in accordance with the predicted scores. After that, the top-ranked 50 candidate miRNAs of the predictive list were checked based on three other prominent databases, dbDEMC [48], miR2Disease [49] and HMDD V3.0 [50]. Finally, we acquired the total frequency of the miRNAs confirmed by one of the three databases or more.

Breast neoplasms are a kind of frequently occurring female carcinoma that leads to high mortality [74]. The treatment is currently-based primarily on the clinical and pathological...
features of breast cancer. Targeted therapy and personalized therapy are the ultimate goals, which are highly dependent on specific molecular typing [75]. For example, the abnormal increase of miR-22 may promote the occurrence and metastasis of breast cancer and lead to a higher degree of malignancy of the tumor [76]. After the implementation of HRNMFMDA, all of the top-ranked 20 and 49 of the top-ranked 50 discovered miRNAs were confirmed by more than one external database (see Table 4).

Colon neoplasms are one of the most high-frequency malignancies of the digestive tract today [77]. According to statistics, in 2019, approximately 101,420 new cases and 51,020 deaths would occur in the United States [78]. Clinical experimental evidence supported that several miRNAs (miR-19a, 98, 146b, 186, 331-5p, 625) were significantly different in patients with rectal adenoma compared to the control group, which could be regarded as potential biomarkers for screening tests [79]. Overall, implementing HRNMFMDA on colon neoplasms confirmed 19 of the top-ranked 20 and 49 out of the top-ranked 50 predicted results were verified through one of the external databases or more (see Supplementary Table S1).

IV. DISCUSSIONS
In our paper, we proposed the Hessian Regularized NMF Method (HRNMFMDA) for revealing uncovered disease-associated miRNAs. In this model, the experimentally discovered miRNA-disease associations, the integrated similarities of miRNAs, and the integrated similarities of disease were merged by the Hessian regularized NMF framework with group LASSO and approximate orthogonal constraints. The proposed model achieved AUC values of 0.9074 and 0.8618 in global and local LOOCV, respectively, which were in 2019, 17,650 esophageal cancer cases will be newly diagnosed, of which 16,080 are deaths in the United States [78]. Several miRNAs have been discovered as biomarkers of esophageal neoplasms. For instance, miR-29c was downregulated in tumor and serum samples from esophageal neoplasm patients, which could be regarded as a promising non-invasive biomarker [81]. After implementing HRNMFMDA on esophageal neoplasms, all of the top-ranked 20 and 49 out of the top-ranked 50 predicted results were verified through one of the external databases or more (see Supplementary Table S2).

### TABLE 4. Top 50 predicted results of breast neoplasms-related miRNAs based on the HMDD V2.0 database.

| miRNA    | Evidence | miRNA    | Evidence |
|----------|----------|----------|----------|
| has-mir-142 | I III | has-mir-181c | I III |
| has-mir-99a | I III | has-mir-30e | I III |
| has-mir-150 | I III | has-mir-95 | I |
| has-mir-106a | I III | has-mir-370 | I III |
| has-mir-192 | I III | has-mir-542 | I III |
| has-mir-138 | I III | has-mir-181d | I II |
| has-mir-15b | I III | has-mir-491 | I III |
| has-mir-130a | I III | has-mir-494 | I III |
| has-mir-378a | I III | has-mir-424 | I III |
| has-mir-196b | I III | has-mir-483 | I |
| has-mir-130b | I III | has-mir-211 | I |
| has-mir-99b | I | has-mir-330 | I |
| has-mir-98 | I I III | has-mir-363 | I III |
| has-mir-92a | I III | has-mir-134 | I III |
| has-mir-186 | I | has-mir-381 | I III |
| has-mir-449a | I III | has-mir-376a | I III |
| has-mir-212 | I III | has-mir-184 | I III |
| has-mir-574 | I III | has-mir-371a | Unconfirmed |
| has-mir-185 | I III | has-mir-432 | I |
| has-mir-144 | I III | has-mir-216a | I |
| has-mir-449b | I | has-mir-650 | I |
| has-mir-32 | I III | has-mir-362 | I |
| has-mir-372 | I III | has-mir-302f | I III |
| has-mir-28 | I | has-mir-198 | I III |
| has-mir-208a | I III | has-mir-517a | I |

Roman numerals I, II, and III represent three databases of dbDEMC, miR2Disease, and HMDD V3.0, respectively.
superior to those of several methods previously reported in the literature. As far as we know, HRNMFMMDA is one of the rare models whose AUC of global LOOCV was higher than 0.9. Furthermore, the model achieved 0.9044+/−0.0080 in 5-fold cross-validation. Moreover, a considerable part of the highest-rated potential related miRNAs was supported by external biological datasets in the case studies. In conclusion, the model can integrate multiple biological data through Hessian regularized NMF effectively for identifying uncovered miRNA-disease pairs.

Reasons for the high predictive capability of the proposed model can be summarized as follows: First, as a semi-supervised method, HRNMFMMDA was not built upon negative samples. Therefore, it may reduce errors caused by generating negative samples randomly. Second, it utilizes the $l_2,1$-norm constraint and approximate orthogonal constraint to select the essential features and ensure the group sparsity of the coding matrix [65], so the influence of noise data can be weakened to obtain relatively reliable prediction results. Third, it can capture the intrinsic manifold structure of data better by introducing Hessian regularization in the NMF framework because the second-order Hessian energy can make the function value change linearly with the geodesic distance [60], which is a universal improvement that deserves more attention.

However, there are still some shortcomings to be solved in our method. The first insufficiency is that the data source we used in HRNMFMMDA may contain noise information. Unfortunately, the error function used in the model was based on the least squares method, which is commonly considered not to be stable enough when dealing with noisy data. Furthermore, it was troublesome to acquire the optimized parameter set of the model. Due to laboratory limitations, we were unable to perform wet experiments to verify the predicted results of the predictive calculation model. Additionally, the calculation of miRNA and disease similarity proposed in the method section of this paper may not be the ideal method. We hope to construct more methods for integrating similarity information more rationally in the future.

ACKNOWLEDGMENT

The authors would like to thank the anonymous peer reviewers for their comments. Moreover, they are grateful to Xing Chen’s Group of China University of Mining and Technology for their guidance and inspiration for this research framework.

REFERENCES

[1] D. P. Bartel, “MicroRNAs: Genomics, biogenesis, mechanism, and function,” Cell, vol. 116, pp. 97–281, Jan. 23 2004.
[2] J. P. O. Correa, E. M. Silva, and F. T. S. Nogueira, “Molecular control by non-coding RNAs during fruit development: From gyosnomci patterning to fruit ripening,” Front Plant Sci, vol. 9, p. 1760, Nov. 2018.
[3] V. Ambros, “The functions of animal microRNAs,” Nature, vol. 431, no. 7006, pp. 350–355, Sep. 2004.
[4] S. Pfeffer, A. Sawyer, M. Lagos-Quintana, R. Sheridan, C. Sander, and F. A. Grasser, “Identification of microRNAs of the herpesvirus family,” Nature Methods, vol. 2, pp. 76–269, Apr. 2005.
[5] D. P. Bartel, “MicroRNAs: Target Recognition and Regulatory Functions,” Cell, vol. 136, no. 2, pp. 215–233, Jan. 2009.
[6] V. K. Gangaaraju and H. Lin, “MicroRNAs: Key regulators of stem cells,” Nature Rev. Mol. Cell Biol., vol. 10, no. 2, pp. 116–125, Feb. 2009.
[7] J. C. Carrington, “Role of MicroRNAs in plant and animal development,” Science, vol. 301, no. 5631, pp. 336–338, Jul. 2003.
[8] S. D. Hatfield, H. R. Shcherbetka, K. A. Fischer, K. Nakahara, R. W. Carthew, and H. Ruohola-Baker, “Stem cell division is regulated by the microRNA pathway,” Nature, vol. 435, no. 7044, pp. 974–978, Jun. 2005.
[9] D. Baumjohann and K. M. Ansel, “MicroRNA-mediated regulation of T helper cell differentiation and plasticity,” Nature Rev. Immunol., vol. 13, no. 9, pp. 666–678, Sep. 2013.
[10] R. A. Boon et al., “MicroRNA-34a regulates cardiac ageing and function,” Nature, vol. 495, no. 7439, pp. 107–110, Mar. 2013.
[11] C. Chebrolu, T. Mostofy, A. Z. Barzilay, and D. Caruana, “A viral Sm-class RNA base-pairs with miRNAs and recruits microRNAs to inhibit apoptosis,” Nature, vol. 550, no. 7675, pp. 275–279, Oct. 2017.
[12] M. Inui, G. Martello, and S. Piccolo, “MicroRNA control of signal transduction,” Nature Rev. Mol. Cell Biol., vol. 11, pp. 63–252, Apr. 2010.
[13] X. Chen, D. Xie, Q. Zhao, and Z.-H. You, “MicroRNAs and complex diseases: From experimental results to computational models,” Briefings Bioinf., vol. 20, no. 2, pp. 515–529, Mar. 2019.
[14] L. Serr et al., “A microRNA-NFATC1 axis links impaired T cell tolerance induction with autoimmune type 1 diabetes,” Sci. Transl. Med., vol. 10, no. 422, Jan. 2018, Art. no. eaag1782.
[15] K. C. Vickers, S. R. Landstreet, M. G. Levin, B. M. Shourci, C. L. Toth, R. C. Taylor, B. T. Palmisano, F. Tabet, H. L. Cui, K.-A. Rye, P. Sethupathy, and A. T. Remaley, “MicroRNA-223 coordinates cholesterol homeostasis,” Proc. Natl. Acad. Sci. USA, vol. 111, no. 44, pp. 14518–14523, Oct. 2014.
[16] S. Moncini, M. Lunghi, A. Valmadre, M. Grasso, V. Del Vescovo, and P. Riva, “The miR-15/107 family of microRNA genes regulates CDK5R1/p35 with implications for Alzheimer’s disease pathogenesis,” Mol. Neurobiol., vol. 54, pp. 4329–4342, Aug. 2017.
[17] K. Jones, J. P. Nourse, C. Keane, A. Bhatnagar, and M. K. Gandhi, “Plasma MicroRNA are disease response biomarkers in classical Hodgkin lymphoma,” Clinical Cancer Res., vol. 20, no. 1, pp. 253–264, Jan. 2014.
[18] R. E. Lanford, E. S. Hildebrandt-Eriksen, A. Petri, R. Persson, L. Lindow, M. E. Munk, S. Kauppinen, and H. Orum, “Therapeutic silencing of MicroRNA-122 in primates with chronic hepatitis C virus infection,” Science, vol. 327, no. 5962, pp. 198–201, Jan. 2010.
[19] S. J. Wheelan, F. Martínez-Murillo, R. A. Irizarry, and J. D. Boeke, “Stacking the deck: Double-tiled DNA microarrays,” Nature Methods, vol. 3, no. 11, pp. 903–907, Nov. 2006.
[20] S. Silva, M. H. S. Paiva, D. R. D. Guedes, L. Krokovsky, F. L. Melo, and M. Silva, “Development and validation of reverse transcription loop-mediated isothermal amplification (RT-LAMP) for rapid detection of ZikaV in mosquito samples from Brazil,” Sci. Rep., vol. 9, p. 4494, Mar. 2019.
[21] Y. Mostovoy, M. Levy-Sakin, J. Lam, E. T. Lam, A. R. Hastie, P. Marks, J. Lee, C. Chu, C. Lin, Z. Džakula, H. Cao, S. A. Schlebusch, K. Giorda, M. Schnall-Levin, J. D. Wall, and P.-Y. Kwok, “A hybrid approach for de novo human genome sequence assembly and phasing,” Nature Methods, vol. 13, no. 7, pp. 587–590, Jul. 2016.
[22] M. Lu, Q. Zhang, M. Deng, J. Miao, Y. Guo, W. Gao, and Q. Cui, “An analysis of human MicroRNA and disease associations,” PLoS ONE, vol. 3, no. 10, p. e3420, Oct. 2008.
[23] D. Wang, J. Wang, M. Lu, F. Song, and Q. Cui, “Infering the human microRNA functional similarity and functional network based on microRNA-associated diseases,” Bioinformatics, vol. 26, no. 13, pp. 1644–1650, Jul. 2010.
[24] Q. Jiang, Y. Hao, G. Wang, L. Juan, T. Zhang, and M. Teng, “Prioritization of disease microRNAs through a human phenotype-microRNAome network,” BMC Syst. Biol., vol. 4, no. Suppl 1, p. s2, May 2010.
[25] Q. Jiang, G. Wang, and Y. Yang, “An approach for prioritizing disease-related microRNAs based on genomic data integration,” in Proc. 3rd Int. Conf. Biomed. Eng. Informat., Oct 2010, pp. 2270–2274.
[26] X. Chen, M.-X. Liu, and G.-Y. Yan, “RWRMDA: Predicting novel human microRNA-disease associations,” Mol. Biosyst., vol. 8, no. 10, p. 2792, 2012.
[27] P. Yuan, K. Han, M. Guo, Y. Guo, J. Li, J. Ding, Y. Liu, Q. Dai, J. Li, Z. Teng, and Y. Huang, “Prediction of microRNAs associated with human diseases based on weighted k most similar neighbors,” PLoS ONE, vol. 8, no. 8, Aug. 2013, Art. no. e70204.
[32] Z.-H. You, Z.-A. Huang, Z. Zhu, G.-Y. Yan, Z.-W. Li, Z. Wen, and X. Chen, "BNPMDA: Bipartite network projection for MiRNA-disease association prediction," Bioinformatics, vol. 34, no. 18, pp. 3178–3186, Sep. 2018.

[33] X. Chen, Z. Zhou, and Y. Zhao, "ELLPMDA: Ensemble learning and link prediction for miRNA-disease association prediction," RNA Biol., vol. 15, pp. 807–818, Jun. 2018.

[34] X. Chen, J.-Y. Cheng, and J. Yin, "Predicting miRNA-disease associations using bipartite local models and hubness-aware regression," RNA Biol., vol. 15, no. 9, pp. 1192–1205, Sep. 2018.

[35] J. Xu, C.-X. Li, Y.-J. Lv, Y.-S. Li, Y. Xiao, T.-T. Shao, H. Liu, X. Li, Y. Zou, Q.-L. Han, X. Li, L.-H. Wang, and H. Ren, "Prioritizing candidate disease miRNAs by topological features in the miRNA target-dysregulated network: Case study of prostate cancer," Mol. Cancer Therapeutics, vol. 10, no. 10, pp. 1857–1866, Oct. 2011.

[36] X. Chen and G. Y. Yan, "Novel semi-supervised learning for potential human MicroRNA-disease associations inference," Sci. Rep., vol. 4, p. 5501, Jun. 2014.

[37] X. Chen, C.-C. Yan, X. Zhang, Z. Li, L. Deng, and Y. Zhang, "RBMMMDA: Predicting multiple types of disease-miRNA associations," Sci. Rep., vol. 5, p. 13877, Sep. 2015.

[38] C. Liang, S. Yu, K. C. Wong, and J. Luo, "A novel semi-supervised model for miRNA-disease association prediction based on L1-norm graph," J. Transl. Med., vol. 16, p. 357, Dec. 2018.

[39] X. Chen, L. Huang, D. Xie, and Q. Zhao, "EGBMMDA: Extreme gradient boosting machine for MiRNA-disease association prediction," Cell Death Disease, vol. 9, no. 1, p. 3, 2018.

[40] X. Chen, L. Wang, J. Qu, N. N. Guan, and J. Q. Li, "Predicting miRNA-disease association based on inductive matrix completion," Bioinformatics, vol. 34, pp. 4256–4265, Dec. 2018.

[41] X. Chen, C.-C. Zhu, and J. Yin, "Ensemble of decision tree reveals potential miRNA-disease associations," PLoS Comput. Biol., vol. 15, no. 7, Jul. 2019, Art. no. e1007209.

[42] X. Chen and L. Huang, "LRSSLMDA: Laplacian Regularized sparse subspace learning for MiRNA-disease association prediction," PLoS Comput. Biol., vol. 13, no. 12, Dec. 2017, Art. no. e1005912.

[43] Y. Zhao, X. Chen, and J. Yin, "Adaptive boosting-based computational model for predicting potential miRNA-disease associations," Bioinformatics, vol. 36, no. 1, p. 330, Jan. 2020.

[44] L. Wang, Z.-H. You, X. Chen, Y.-M. Li, Y.-N. Dong, L.-P. Li, and K. Zheng, "LMTRDA: Using logistic model tree to predict MiRNA-disease associations by fusing multi-source information of sequences and similarities," PLoS Comput. Biol., vol. 15, no. 3, Mar. 2019, Art. no. e1006865.

[45] X. Chen, C.-C. Wang, J. Yin, and Z.-H. You, "Novel human miRNA-disease association inference based on random forest," Mol. Therapy Nucleic Acids, vol. 13, pp. 568–579, Dec. 2018.

[46] C.-C. Wang, X. Chen, J. Yin, and J. Qu, "An integrated framework for the identification of potential miRNA-disease association based on negative samples extraction strategy," RNA Biol., vol. 16, no. 3, pp. 257–269, Mar. 2019.

[47] L. Zhang, X. Chen, and J. Yin, "Prediction of potential miRNA-disease associations through a novel unsupervised deep learning framework with variational autoencoder," Cells, vol. 8, p. 1040, Sep. 2019.

[48] Z. Fang, F. Ren, L. Sun, S. He, G. Sun, G. Qiao, L. Yao, Y. Zhang, R. Miao, Y. Cao, Y. Zhao, Y. Zhong, and H. Zhao, "DdDEMCA: A database of differentially expressed miRNAs in human cancers," BMC Genomics, vol. 11, no. Suppl 4, p. 55, Dec. 2010.

[49] Q. Jiang, Y. Wang, Y. Hao, L. Juan, M. Teng, X. Zhang, M. Li, G. Wang, and Y. Liu, "MiRD2ADE: A manually curated database for microRNA deregulation in human disease," Nucleic Acids Res., vol. 37, pp. D98–D104, Jun. 2009.
[75] D. Zardavas, J. Baselga, and M. Piccart, “Emerging targeted agents in metastatic breast cancer,” Nature Rev. Clin. Oncol., vol. 10, no. 4, pp. 191–210, Apr. 2013.

[76] P. Dalerba and M. Clarke, “Oncogenic miRNAs and the perils of losing control of a stem cell’s epigenetic identity,” Cell Stem Cell, vol. 13, no. 1, pp. 5–6, Jul. 2013.

[77] P. Favoriti, G. Carbone, M. Greco, F. Pirozzi, R. E. M. Pirozzi, and F. Corcione, “Worldwide burden of colorectal cancer: A review,” Updates Surg., vol. 68, no. 1, pp. 7–11, Mar. 2016.

[78] R. L. Siegel, K. D. Miller, and A. Jemal, “Cancer statistics, 2019,” CA Cancer J. Clin., vol. 69, pp. 7–34, Jan. 2019.

[79] A. M. Verma, M. Patel, M. I. Aslam, J. Jameson, J. H. Pringle, P. Wurm, and B. Singh, “Circulating plasma microRNAs as a screening method for detection of colorectal adenomas,” Lancet, vol. 385, p. $100, Feb. 2015.

[80] E. P. Simard, E. M. Ward, R. Siegel, and A. Jemal, “Cancers with increasing incidence trends in the United States: 1999 through 2008,” CA, Cancer J. Clinicians, vol. 62, no. 2, pp. 118–128, Mar. 2012.

[81] B. Li 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, vol. 9, no. 6, pp. 1599–1613, 2019.

**YONG TANG** is currently pursuing the Ph.D. degree with the Department of Health Statistics, College of Preventive Medicine, Army Medical University. His research interest includes computational biology and bioinformatics.

**GAOMING LI** is currently pursuing the Ph.D. degree with the Department of Health Statistics, Army Medical University, China. His research interests include bioinformatics and biostatistics.

**YAZHOU WU** received the Ph.D. degree in biomedical engineering from Third Military Medical University, China, in 2007. He is currently a Professor with the Department of Health Statistics, College of Preventive Medicine, Army Medical University. His research interests include bioinformatics, multiomics data analysis based on deep learning method and biostatistics.

**DONG YI** received the Ph.D. degree in computer science from Chongqing University, China, in 1997. He is currently a Professor with the Department of Health Statistics, College of Preventive Medicine, Army Medical University. His research interests include bioinformatics and medical statistics.

* * *