Front. Comput. Sci., 2024, 18(2): 182903
https://doi.org/10.1007/s11704-023-2490-5

RESEARCH ARTICLE

# Identification of human microRNA-disease association via low-rank approximation-based link propagation and multiple kernel learning

**Yizheng WANG[1,2,*], Xin ZHANG[3,*], Ying JU[4], Qing LIU[5], Quan ZOU[1,2], Yazhou ZHANG[6], Yijie DING (✉)[2], Ying ZHANG (✉)[5]**

1 Institute of Fundamental and Frontier Sciences, University of Electronic Science and Technology of China, Chengdu 610054 , China
2 Yangtze Delta Region Institute (Quzhou), University of Electronic Science and Technology of China, Quzhou 324000, China
3 Beidahuang Industry Group General Hospital, Harbin 150088, China
4 School of Informatics, Xiamen University, Xiamen 361005, China
5 Department of Anesthesiology, Hospital (T.C.M) Affiliated to Southwest Medical University, Luzhou 646000, China
6 Software Engineering College, Zhengzhou University of Light Industry, Zhengzhou 450002, China



**Abstract** Numerous studies have demonstrated that human microRNAs (miRNAs) and diseases are associated and studies on the microRNA-disease association (MDA) have been conducted. We developed a model using a low-rank approximation-based link propagation algorithm with Hilbert–Schmidt independence criterion-based multiple kernel learning (HSIC-MKL) to solve the problem of the large time commitment and cost of traditional biological experiments involving miRNAs and diseases, and improve the model effect. We constructed three kernels in miRNA and disease space and conducted kernel fusion using HSIC-MKL. Link propagation uses matrix factorization and matrix approximation to effectively reduce computation and time costs. The results of the experiment show that the approach we proposed has a good effect, and, in some respects, exceeds what existing models can do.

**Keywords** human miRNA-disease association, multiple kernel learning, link propagation, miRNA similarity, disease similarity

## 1 Introduction

Human miRNAs are noncoding RNAs and play several important regulatory roles. Typically, they contain between 20 and 25 nucleotides [1]. They are closely related to life and other biological processes, including cell differentiation, development, and proliferation [2,3]. Studies have found that the expression levels of miRNAs are associated with human diseases such as Lymphoma, Alzheimer's disease and Diabetes [4−10]. For example, previous research has shown that hsa-miR-146a is associated with human breast cancer, and the metastasis of cancer cells can be controlled by its expression levels [11−14]. Therefore, the identification of MDA is helpful to explore the causes of human diseases and is crucial for the diagnosis and treatment of human diseases [15,16]. Traditional biological experimental methods involve complex experimental steps and expensive experimental instruments, and are time-consuming, and have a high experimental cost. Hence, the need to develop computational techniques for accurately and swiftly identifying MDA is critical.

In recent years, researchers have successfully created several computational approaches for identifying MDA [17−19]. Network-based approaches [20] and machine learning-based approaches [21,22] can be used to categorize these studies. Constructing feature kernels and using them to build appropriate models to predict potential associations are the two most important steps in network-based approaches. To identify MDA, many previous studies used machine learning-based approaches [23−25]. By using a variety of information to construct kernels, such as the functional kernel and known MDA in miRNA space and the semantic kernel of disease and Gaussian Interaction Profile (GIP) kernel in disease space, Chen et al. [26] created the WBSMDA model for identifying MDA. Chen et al. [27] developed the MDHGI model by performing matrix decomposition before using heterogeneous graphs to inference MDA prediction. To identify MDA, Ding et al. [28] utilized a hypergraph regularized bipartite local model in combination with Laplace SVM. Chen et al. [29]

Received July 29, 2022; accepted February 7, 2023

E-mail: wuxi_dyj@163.com; zhangying021210@163.com

* These authors contributed equally to this work.

developed a model for predicting MDA using inductive matrix completion, called IMCMDA. In the model NCMCMDA [30], neighborhood constraint matrix completion is used to predict MDA. Neural networks have also been applied in previous studies. A deep ensemble model, the DeepMDA proposed by Fu et al. [31], extracts features from similarity information and makes predictions using a three-layer neural network. The PCFM model proposed by Zeng et al. [32−34] is a probabilistic collaborative filtering model that effectively reduces the computational cost but also achieves good performance. Additionally, the development of the model has involved a great deal of work from numerous researchers. To determine potential MDA, Chen et al. [35] used random walk with restart for MDA (RWRMDA). Van Laarhoven et al. [36] utilized the GIP similarity kernel to raise the effectiveness of WBSMDA. Gu et al. [37] created a Network Consistency Projection for the MDA (NCPMDA) approach to discover the potential MDA.

Our work aims to construct a well-behaved model for predicting MDA by utilizing the enormous amount of genomic data presently available. We first constructed three similarity kernels in the feature space, miRNA, and disease, and comprehensively considered their GIP. In addition, sequence information and functional information of miRNAs, as well as semantic and functional information of diseases, are also used. By using HSIC-MKL, the dependence of the feature space and label space are maximized to obtain the weights used to fuse the kernels. By using the normalized fuse kernels and the known association information, low-rank approximation-based link propagation is used to make association predictions.

In this paper, we document a number of significant contributions made by our research to the field of bioinformatics: (1) HSIC-MKL is used to effectively fuse the similarity kernels of miRNA space and disease space; (2) The low-rank approximation-based link propagation is used, which significantly reduces the computational cost while maintaining good accuracy; (3) Experiments show that our model achieves excellent results, and outperforms current techniques.

## 2 Materials and methods

In our study, we established similarity kernels of miRNA and disease to predict MDA, and the kernel methods has been applied in various previous studies [38]. HSIC-MKL was used to fuse kernels. Then, the fused kernels were normalized. By using the similarity information of links and nodes, link propagation based on matrix decomposition is used to make association predictions. A schematic diagram of our approach is shown in Fig. 1.

### 2.1 Problem description

There are two node sets $M \equiv \{m_1, m_2, \ldots, m_p\}$ and $D \equiv \{d_1, d_2, \ldots, d_q\}$ in the benchmark dataset that can be downloaded from the HMDD database [39], representing miRNA and disease, respectively. In this paper, $p \equiv |M|$ and $q \equiv |D|$ are defined as the number of nodes, and their values are 495 and 383, respectively. Adjacency matrix $\boldsymbol{Y}$ stores the 5430 associations between miRNA and disease and satisfies $\boldsymbol{Y} \in \boldsymbol{R}^{p \times q}$, and the elements therein are defined as

$$[Y]_{i,j} \equiv \begin{cases} 1, \text{if } m_i \text{ associated with } d_j, \\ 0, \text{otherwise.} \end{cases} \quad (1)$$

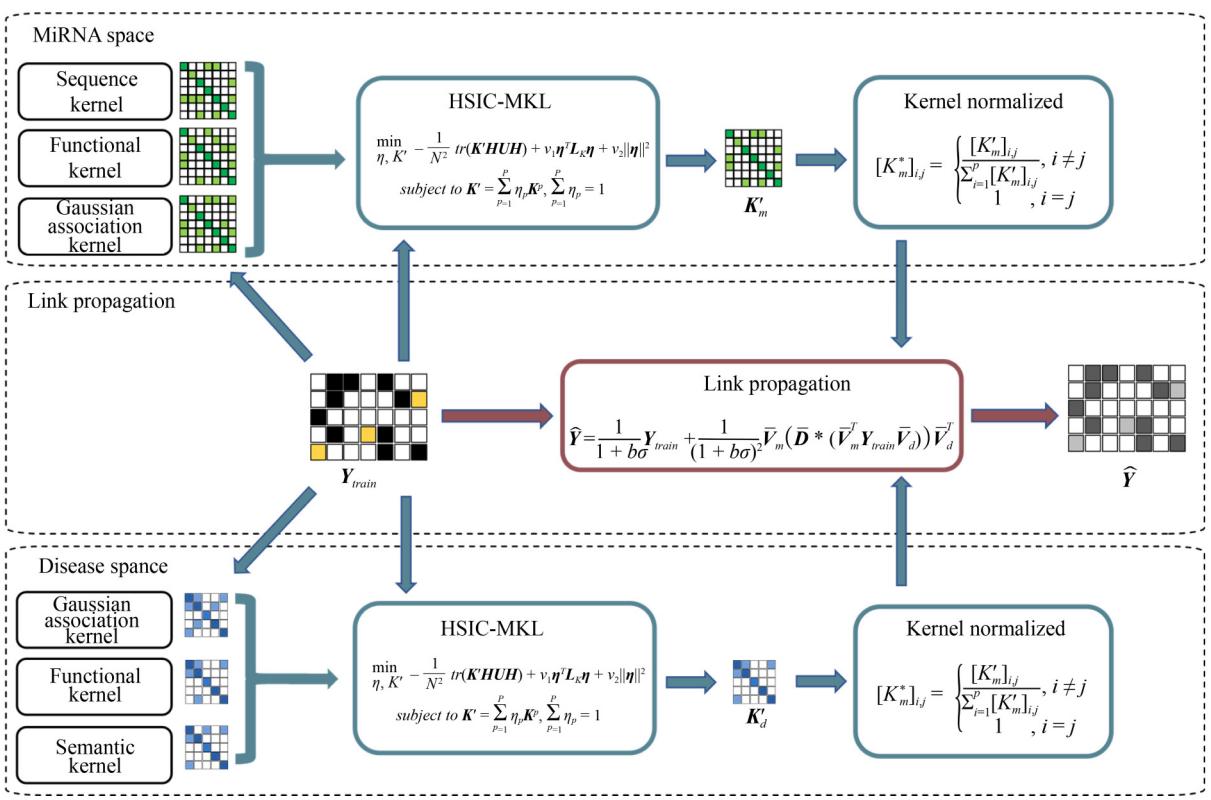



**Fig. 1** Schematic diagram of our proposed method

Our aim was to predict the possibility of an association being established between two nodes. We call this possibility link strength and store it in a matrix $\widehat{\boldsymbol{Y}}$ of the same size as $\boldsymbol{Y}$. The larger the element $\widehat{Y}_{i,j}$ in the matrix $\widehat{\boldsymbol{Y}}$ is, the more confident the algorithm is about the association between $m_i$ and $d_j$.

To complete the semi-supervised learning task and better evaluate the performance, we randomly mask the adjacency matrix $\boldsymbol{Y}$. The process of masking is shown in Fig. 2. The black square represents the association while white square represents the non-association between miRNA and disease in the corresponding position of the matrix. Then, we obtain the masked matrix $\boldsymbol{Y}_{train}$ by randomly changing some elements of the matrix $\boldsymbol{Y}$ to 0, where the masked parts are represented by yellow squares and regarded as the validation set or testing set. And the unmasked parts are treated as the training set. and the rest is the training set. The algorithm calculates the matrix $\widehat{\boldsymbol{Y}}$ by predicting the MDA. In the matrix $\widehat{\boldsymbol{Y}}$, the masked part is represented by gray square.

In this study, we employed three techniques to assess the model's efficacy. We used 5-fold cross validation (5-CV) to assess the performance of predicting the link strength of an association between miRNA and disease on the adjacency matrix $\boldsymbol{Y}$. In 5-CV, all MDA are randomly divided into 5 groups. One group serves as a test set, while the remaining four groups serve as training sets. In the global leave-one-out cross validation (global LOOCV), every one of the 5430 known MDA is taken into consideration as the test samples for research, while the rest are handled as training samples. In particular, 5-CV and global LOOCV evaluate the model's ability to exploit existing MDA to discover potential MDA. In biology, there is a situation where if a miRNA is newly discovered and does not have any known MDA, its association with disease needs to be predicted by the model. And then local leave-one-out cross validation (local LOOCV) is used to evaluate the model's performance of predicting the associations of newly discovered miRNAs with diseases. The local LOOCV only considers the disease association of a specific miRNA (corresponding to a column of the matrix $\boldsymbol{Y}$). In the local LOOCV method, one row of the matrix $\boldsymbol{Y}$ is completely masked as a test set and the rest as a training set, which is shown in Fig. 2.

## 2.2 MiRNA kernels

There are three similarity matrices used to describe the similarity of miRNAs: the sequence similarity kernel, the functional similarity kernel, and the GIP kernel, which are widely used in related previous studies [40,41].

**MiRNA sequence kernel**

From the miRBase [42] database, we obtained all 495 miRNA sequences. The similarity of the sequence can then be extracted using the Needleman-Wunsch algorithm. After that, the information is kept in the kernel matrix $\boldsymbol{K}_{m,1} \in \boldsymbol{R}^{p\times p}$.

**MiRNA functional kernel**

In accordance with the MISIM proposed by Wang et al. [43], we created a miRNA functional similarity kernel $\boldsymbol{K}_{m,2} \in \boldsymbol{R}^{p\times p}$. The disease semantic similarity, and established links between miRNAs and diseases were used in this strategy to structure the kernel of miRNA functional similarity $\boldsymbol{K}_{m,2}$.

**Gaussian interaction profile kernel for miRNAs**

Fig. 3 shows the calculation process of the GIP kernel. For all associations of miRNA and disease with random masks, we obtain the adjacency matrix $\boldsymbol{Y}_{train}$, where 1 represents the association and 0 represents either non-association or masked samples. In Fig. 3, $\boldsymbol{pr}_{m_3}$ is a disease profile that corresponds to a specific miRNA $m_3$. All 495 disease profiles of miRNAs can be obtained from the adjacency matrix $\boldsymbol{Y}_{train}$. Then, the elements of the GIP kernel matrix can be calculated as

$$\boldsymbol{K}_{m,3}\left(m_i, m_j\right) = \exp\left(-\gamma\left\|\boldsymbol{pr}_{m_i} - \boldsymbol{pr}_{m_j}\right\|^2\right), \tag{2}$$

where $\boldsymbol{pr}_{m_i}$ and $\boldsymbol{pr}_{m_j}$ are the disease profiles of miRNAs $m_i$ and $m_j$, respectively, and $\gamma$ denotes the bandwidth and is set to 1 in this paper.

## 2.3 Disease kernels

In our study, we use the semantic similarity kernel, the

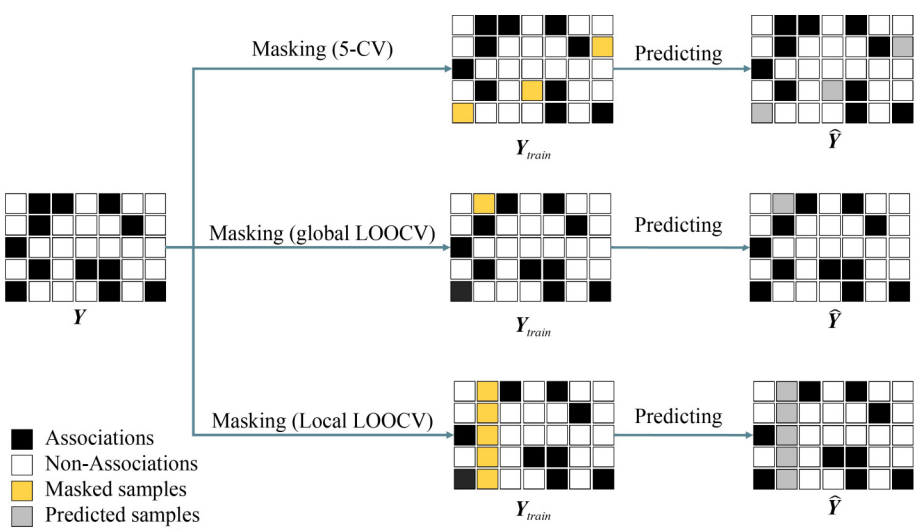



**Fig. 2** Schematic of training and testing

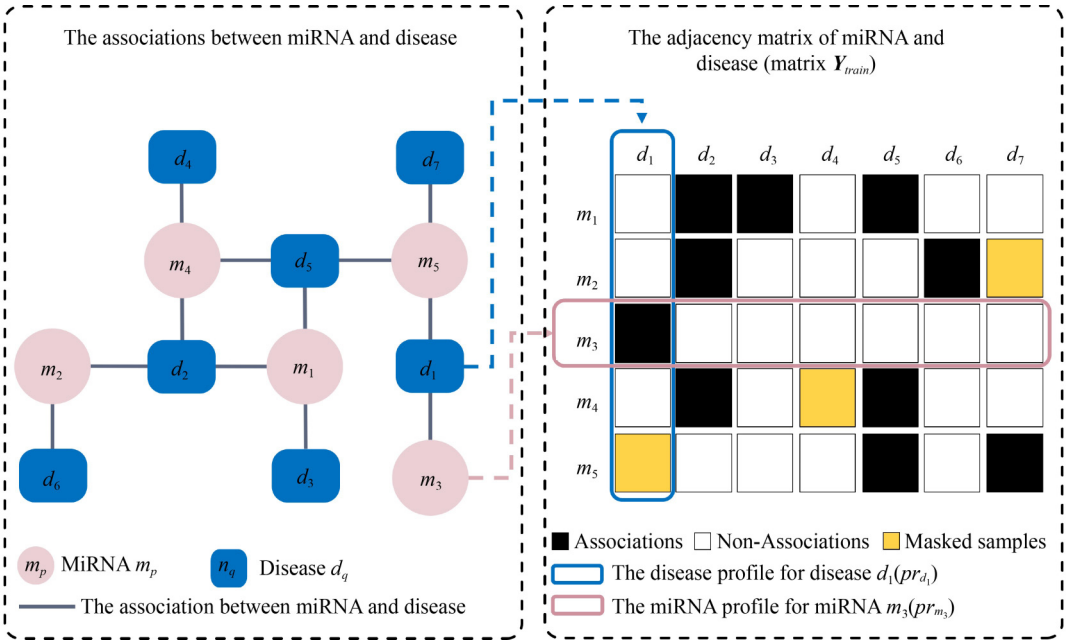



**Fig. 3** The profile of miRNA and disease

functional similarity kernel, and the GIP kernel, to describe disease information, and the effectiveness of using these kernels has been widely demonstrated in previous studies [44,45].

**Disease semantic kernel**

In the MeSH [46] database, directed acyclic graphs (DAG) are used to represent disease information. In DAG, diseases are treated as nodes, and edges represent connections between diseases. We define the set of ancestor nodes of a particular disease $d_i$ as $P_{d_i}$ and the set of corresponding edges as $E_{d_i}$. Then, the set $T_{d_i}$ is defined as follows:

$$T_{d_i} \equiv P_{d_i} \cap \{d_i\}. \tag{3}$$

DAG can be defined as follows:

$$DAG_{d_i} = \left(d_i, T_{d_i}, E_{d_i}\right), \tag{4}$$

and then the semantic score for disease $t \in T_{d_i}$ is

$$D_{d_i}(t) = \begin{cases} 1 & ,\text{if } t = d_i, \\ \max\left\{\rho * D_{d_i}\left(t'\right) | t' \in children\ of\ t\right\} & ,\text{if } t \neq d_i, \end{cases} \tag{5}$$

where $\rho$ is the contribution factor, and its value is 0.5 in our study.

After that, the semantic score of the disease $d_i$ is defined as

$$DV(d_i) = \sum\nolimits_{t \in T_{d_i}} D_{d_i}(t), \tag{6}$$

we can calculate the elements of semantic kernel matrix $\boldsymbol{K}_{d,1} \in \boldsymbol{R}^{q \times q}$ as

$$\boldsymbol{K}_{d,1}\left(d_i, d_j\right) = \frac{\sum_{t \in T_{d_i} \cap T_{d_j}} \left(D_{d_i}(t) + D_{d_j}(t)\right)}{DV(d_i) + DV\left(d_j\right)}. \tag{7}$$

**Disease functional kernel**

Calculating the functional similarity of diseases involves associations between genes and diseases [47,48]. The log likelihood score (LLS) can be obtained from the HumanNet database [49], which measures the probability that two genes in the database have a functional relationship. The normalized LLS can be computed as

$$L^*(g_k, g_s) = \frac{L(g_k, g_s) - L_{min}}{L_{max} - L_{min}}, \tag{8}$$

where $L(g_k, g_s)$ and $L^*(g_k, g_s)$ are the LLS and normalized LLS between the $k$th and $s$th genes, respectively. In HumanNet, $L_{min}$ and $L_{max}$ represent the minimum and maximum values of LLS, respectively.

In this study, $FS(g_k, g_s)$ is used to represent the functional similarity score between the $k$th and $s$-th genes, and its equation is as follows:

$$FS(g_k, g_s) = \begin{cases} 1 & ,\text{if } k = s, \\ L^*(g_k, g_s) & ,\text{if } k \neq s \cap e(k,s) \in S_{HumanNET}, \\ 0 & ,\text{if } k \neq s \cap e(k,s) \notin S_{HumanNET}, \end{cases} \tag{9}$$

where $e(k,s)$ represents the association between two genes and $S_{HumanNET}$ is the set containing all genetic associations in the database HumanNet.

Then, the functional similarity score between a gene $g$ and a set of genes $G$ is as follows:

$$F_G(g) = \max_{g_s \in G} FS(g, g_s). \tag{10}$$

We can download the association between diseases and genes in SIDD [50]. Diseases $d_i$ and $d_j$ correspond to gene sets $G_i$ and $G_j$, respectively, and then we can calculate the elements of the functional similarity score $\boldsymbol{K}_{d,2} \in \boldsymbol{R}^{q \times q}$ as

$$\boldsymbol{K}_{d,2}\left(d_i, d_j\right) = \frac{\sum_{g_k \in G_j} F_{G_i}(g_k) + \sum_{g_s \in G_i} F_{G_j}(g_s)}{|G_i| + \left|G_j\right|}. \tag{11}$$

**Gaussian interaction profiles kernel for diseases**

Similarly, we obtained the miRNA profile of all diseases in the adjacency matrix Y, so that each element of the GIP kernel

matrix $\boldsymbol{K}_{d,3} \in \boldsymbol{R}^{q\times q}$ can be computed as

$$\boldsymbol{K}_{d,3}\left(d_i,d_j\right)=\exp\left(-\gamma\left\|\boldsymbol{pr}_{d_i}-\boldsymbol{pr}_{d_j}\right\|^2\right), \tag{12}$$

where the value of $\gamma$ is 1.

### 2.4 Hilbert–Schmidt Independence Criterion-based multiple kernel learning

As shown in Table 1, we acquired three miRNA kernels and three disease kernels. We can calculate the weight of miRNA kernels and disease kernels through multiple kernel learning (MKL) to obtain the optimal kernel. In this way, we can use all the kernels of miRNA space and disease space, and the model adopts more information, which aids in enhancing the performance of the model. By defining a kernel set $\boldsymbol{K}$ containing $P$ kernels, the optimal kernel can be obtained as

$$\boldsymbol{K}'=\sum_{p=1}^{P}\boldsymbol{\eta}_p\boldsymbol{K}^p, \tag{13}$$

where $\boldsymbol{K}^p \in \boldsymbol{R}^{N\times N}$, and $\boldsymbol{\eta}_p$ denotes the weight of the pth kernel. To calculate the weight, HSIC-MKL [51,52] is employed.

We define that $\boldsymbol{X}=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\ldots,\boldsymbol{x}_N\}^T \in \boldsymbol{R}^{N\times d}$ is the original feature of $d$ dimensions of $N$ samples, and $\boldsymbol{Y} \in \boldsymbol{R}^{N\times 1}$ is the labels of these samples. We can derive a series of observations from probability distribution $Pr_{xy}$, defined as

$$\boldsymbol{Z} \equiv \{(\boldsymbol{x}_1,y_1),\ldots,(\boldsymbol{x}_N,y_N)\} \subseteq \boldsymbol{X}\times\boldsymbol{Y}. \tag{14}$$

HSIC calculates the cross-covariance operator on the domain $\boldsymbol{X}\times\boldsymbol{Y}$ to determine the independence between $\boldsymbol{X}$ and $\boldsymbol{Y}$. The feature set $\boldsymbol{X}$ and label set $\boldsymbol{Y}$ can be mapped to $\boldsymbol{F}$ and $\boldsymbol{G}$ by the mapping $\phi:\boldsymbol{X}\to\boldsymbol{F}$ and $\psi:\boldsymbol{Y}\to\boldsymbol{G}$. Then, we defined their expectations as $\mu_{\boldsymbol{x}}$ and $\mu_y$, respectively. The kernel function of $\boldsymbol{X}$ is as follows:

$$k\left(\boldsymbol{x}_i,\boldsymbol{x}_j\right)=\left\langle\phi(\boldsymbol{x}_i),\phi\left(\boldsymbol{x}_j\right)\right\rangle. \tag{15}$$

Similarly, the kernel function of $\boldsymbol{Y}$ is defined as

$$l\left(y_i,y_j\right)=\left\langle\psi(y_i),\psi\left(y_j\right)\right\rangle. \tag{16}$$

The following equation can be used to determine the cross-covariance operator $C_{\boldsymbol{xy}}$:

$$C_{\boldsymbol{xy}}=E_{\boldsymbol{x},y}\left[\phi(\boldsymbol{x})\otimes\psi(y)\right]-\mu_{\boldsymbol{x}}\mu_y, \tag{17}$$

where $E_{\boldsymbol{x},y}$ denotes the common expectation of $\boldsymbol{x}$ and $y$. Then, we can write the HSIC operator as:

$$HSIC\left(\boldsymbol{F},\boldsymbol{G},Pr_{xy}\right)=\left\|C_{\boldsymbol{xy}}\right\|_{HS}^2. \tag{18}$$

Then, we define $\boldsymbol{I}$ as the identity matrix, and it satisfies $\boldsymbol{I} \in \boldsymbol{R}^{N\times N}$. By defining $\boldsymbol{e}=[1,\ldots,1]^T$, we can obtain

**Table 1** Summary of similarities for two feature spaces

| Space | Kernel | Description |
|---|---|---|
| MiRNA | $K_{m,1}$ | Sequence information of miRNA |
| | $K_{m,2}$ | Functional information of miRNA |
| | $K_{m,3}$ | Gaussian interaction profiles for miRNA |
| Disease | $K_{d,1}$ | Semantic information of disease |
| | $K_{d,2}$ | Functional information of disease |
| | $K_{d,3}$ | Gaussian interaction profiles for disease |

$$\boldsymbol{H} \equiv \boldsymbol{I}-\frac{\boldsymbol{ee}^T}{N}. \tag{19}$$

Note that $\boldsymbol{H}$ is the centering matrix, and it satisfies $\boldsymbol{H} \in \boldsymbol{R}^{N\times N}$. Then, we can make an empirical estimate of $\boldsymbol{Z}$ set as

$$\begin{aligned}HSIC(\boldsymbol{F},\boldsymbol{G},\boldsymbol{Z})&=\frac{1}{N^2}tr(\boldsymbol{KU})-\frac{2}{N^3}\boldsymbol{e}^T\boldsymbol{KUe}+\frac{1}{N^4}\boldsymbol{e}^T\boldsymbol{Kee}^T\boldsymbol{Ue}\\&=\frac{1}{N^2}\left[tr(\boldsymbol{KU})-\frac{1}{N}tr\left(\boldsymbol{KUee}^T\right)-\frac{1}{N}tr\left(\boldsymbol{UKee}^T\right)\right.\\&\quad\left.+\frac{1}{N^2}tr\left(\boldsymbol{Uee}^T\boldsymbol{Kee}^T\right)\right]\\&=\frac{1}{N^2}tr\left[\boldsymbol{K}\left(\boldsymbol{I}-\frac{1}{N}\boldsymbol{ee}^T\right)\boldsymbol{U}\left(I-\frac{1}{N}\boldsymbol{ee}^T\right)\right]\\&=\frac{1}{N^2}tr(\boldsymbol{KHUH}) \triangleq HSIC(\boldsymbol{K},\boldsymbol{U}),\end{aligned} \tag{20}$$

where $\boldsymbol{K},\boldsymbol{U} \in \boldsymbol{R}^{N\times N}$ are kernel matrices, as $\boldsymbol{K}_{ij}=k\left(\boldsymbol{x}_i,\boldsymbol{x}_j\right)$ and $\boldsymbol{U}_{ij}=l\left(y_i,y_j\right)$. Note that the value of $HSIC(\boldsymbol{K},\boldsymbol{U})$ is related to the dependence of $\boldsymbol{K}$ and $\boldsymbol{U}$, and the higher it is, the stronger the dependence of $\boldsymbol{K}$ and $\boldsymbol{U}$. In addition, it should be in the range of 0 to 1; when it is equal to 0, we think that $\boldsymbol{K}$ and $\boldsymbol{U}$ are independent or irrelevant. We define Frobenius inner as

$$\langle\boldsymbol{K},\boldsymbol{U}\rangle_F=tr\left(\boldsymbol{K}^T\boldsymbol{U}\right). \tag{21}$$

Then, we defined $\boldsymbol{W}$ as the cosine similarity matrix between two kernels satisfying $\boldsymbol{W} \in \boldsymbol{R}^{P\times P}$, and the equation is as follows:

$$Aligned(\boldsymbol{K},\boldsymbol{U})=\frac{\langle\boldsymbol{K},\boldsymbol{U}\rangle_F}{\|\boldsymbol{K}\|_F\|\boldsymbol{U}\|_F}, \tag{22}$$

where $\|\boldsymbol{K}\|_F=\sqrt{\langle\boldsymbol{K},\boldsymbol{K}\rangle_F}$ is the Frobenius norm. $\boldsymbol{D}_K$ is defined as a diagonal matrix that satisfies $\boldsymbol{W} \in \boldsymbol{R}^{P\times P}$, and its elements can be calculated as

$$[D_K]_{i,j}=\sum_{j=1}^{p}[W]_{i,j}. \tag{23}$$

Then, the graph Laplacian matrix $\boldsymbol{L}_k \in \boldsymbol{R}^{P\times P}$ is defined as

$$\boldsymbol{L}_k=\boldsymbol{D}_k-\boldsymbol{W}, \tag{24}$$

we can write the Laplacian regular term as

$$\begin{aligned}\sum_{i,j}^{P}\left(\eta_i-\eta_j\right)^2W_{ij}&=\sum_{i,j}^{P}\left(\eta_i^2+\eta_j^2-2\eta_i\eta_j\right)W_{ij}\\&=\sum_i^P\eta_i^2D_{ii}+\sum_j^P\eta_j^2D_{jj}-2\sum_{i,j}^{P}\eta_i\eta_jW_{ij}\\&=2\boldsymbol{\eta}^T\boldsymbol{L}_K\boldsymbol{\eta}.\end{aligned} \tag{25}$$

Inspired by multiple kernel learning [53] and HSIC [51,54] and classified by key properties [55], we obtain the convex optimization problem and its objective function as follows:

$$\begin{aligned}&\min_{\eta,K'}-\frac{1}{N^2}tr\left(\boldsymbol{K}'\boldsymbol{HUH}\right)+\nu_1\boldsymbol{\eta}^T\boldsymbol{L}_K\boldsymbol{\eta}+\nu_2\|\boldsymbol{\eta}\|^2,\\&subject\ to\ \boldsymbol{K}'=\sum_{p=1}^{P}\eta_p\boldsymbol{K}^p,\\&\qquad\sum_{p=1}^{P}\eta_p=1,\end{aligned} \tag{26}$$

where the first term comes from HSIC, which measures the dependence between $\boldsymbol{K}'$ and $\boldsymbol{U}$ and makes the label space and feature space more dependent by minimizing its value. The second term is the graph regularization term, which assists in smoothing the weights, and the last term is the $L2$ norm regularization. Note that $\boldsymbol{\eta}$ is the kernel weight, and $\boldsymbol{K}'$ is the optimal kernel. When we use the above function to estimate the weight $\boldsymbol{\eta}_m$ to obtain the optimal miRNA kernel $\boldsymbol{K}'_m$, the label kernel $\boldsymbol{U}$ is written as

$$\boldsymbol{U} = \boldsymbol{Y}_{train}\boldsymbol{Y}_{train}^{T}. \tag{27}$$

For the optimal disease kernel $\boldsymbol{K}'_d$, we can write

$$\boldsymbol{U} = \boldsymbol{Y}_{train}^{T}\boldsymbol{Y}_{train}. \tag{28}$$

Finally, the normalized miRNA kernel $\boldsymbol{K}_m^*$ [56] can be calculated as:

$$[K_m^*]_{i,j} = \begin{cases} 1 & ,i = j, \\ \dfrac{[K'_m]_{i,j}}{\sum_{i=1}^{p}[K'_m]_{i,j}} & ,i \neq j. \end{cases} \tag{29}$$

Similarly, the normalized optimal kernel $\boldsymbol{K}_d^*$ can also be obtained.

### 2.5 Link propagation

As in label propagation [57,58], by changing the application scope of the label propagation principle from a single node to a pair of nodes, the principle of link propagation is expressed as "two nodes similar to each other may have the same link strength" [59]. We obtain the following objective function to minimize:

$$J\left(\widehat{\boldsymbol{Y}}\right) = \frac{1}{2}\left\|vec\left(\widehat{\boldsymbol{Y}}\right) - vec\left(\boldsymbol{Y}_{train}\right)\right\|_2^2 + \frac{\sigma}{2}vec\left(\widehat{\boldsymbol{Y}}\right)^{T} Lvec\left(\widehat{\boldsymbol{Y}}\right), \tag{30}$$

where the first term, which serves as a loss function, comes from the Kronecker Regularized Least Squares approach, which matches the prediction $\widehat{\boldsymbol{Y}}$ and target values $\boldsymbol{Y}_{train}$ of the known part of the network. In addition, it is a regularization term to increase numerical stability and makes the prediction close to zero. The second term indicates that the greater the similarity between two pairs, the closer the value of their link strength values. $\sigma$ plays a role in balancing the first term and the second term, and it should be greater than 0. $\boldsymbol{L}$ is a Laplacian matrix and satisfies $\boldsymbol{L} \in \boldsymbol{R}^{pq\times pq}$.

Laub [60] defined $\otimes$ and $\oplus$to represent the Kronecker product operator and the Kronecker sum operator, respectively. Kashima et al. [61] proposed using the Kronecker product Laplacian or Kronecker sum Laplacian, and the corresponding equations are as follows:

$$\boldsymbol{L} \equiv \boldsymbol{D}_d \otimes \boldsymbol{D}_m - \boldsymbol{K}_d^* \otimes \boldsymbol{K}_m^*, \tag{31}$$

$$\boldsymbol{L} \equiv \boldsymbol{D}_d \oplus \boldsymbol{D}_m - \boldsymbol{K}_d^* \oplus \boldsymbol{K}_m^*, \tag{32}$$

where $\boldsymbol{D}_m$ and $\boldsymbol{D}_d$ are diagonal matrices defined as

$$[D_m]_{i,i} \equiv \sum\nolimits_j [K_m^*]_{i,j}, \tag{33}$$

$$[D_d]_{i,i} \equiv \sum\nolimits_j \left[K_d^*\right]_{i,j}. \tag{34}$$

Moreover, we use the normalized Kronecker product Laplacian matrix and Kronecker sum Laplacian matrix to facilitate derivation of the solution:

$$\begin{aligned} \boldsymbol{L} &\equiv \boldsymbol{I} - (\boldsymbol{D}_d \otimes \boldsymbol{D}_m)^{-\frac{1}{2}}\left(\boldsymbol{K}_d^* \otimes \boldsymbol{K}_m^*\right)(\boldsymbol{D}_d \otimes \boldsymbol{D}_m)^{-\frac{1}{2}} \\ &= \boldsymbol{I} - \left(\boldsymbol{D}_d^{-\frac{1}{2}}\boldsymbol{K}_d^*\boldsymbol{D}_d^{-\frac{1}{2}}\right) \otimes \left(\boldsymbol{D}_m^{-\frac{1}{2}}\boldsymbol{K}_m^*\boldsymbol{D}_m^{-\frac{1}{2}}\right), \end{aligned} \tag{35}$$

$$\begin{aligned} \boldsymbol{L} &\equiv \boldsymbol{I} - \left(\boldsymbol{I} - \boldsymbol{D}_d^{-\frac{1}{2}}\boldsymbol{K}_d^*\boldsymbol{D}_d^{-\frac{1}{2}}\right) \oplus \left(\boldsymbol{I} - \boldsymbol{D}_m^{-\frac{1}{2}}\boldsymbol{K}_m^*\boldsymbol{D}_m^{-\frac{1}{2}}\right) \\ &= 3\boldsymbol{I} - \left(\boldsymbol{D}_d^{-\frac{1}{2}}\boldsymbol{K}_d^*\boldsymbol{D}_d^{-\frac{1}{2}}\right) \oplus \left(\boldsymbol{D}_m^{-\frac{1}{2}}\boldsymbol{K}_m^*\boldsymbol{D}_m^{-\frac{1}{2}}\right). \end{aligned} \tag{36}$$

As a matter of convenience, we define the following matrices:

$$\widetilde{\boldsymbol{K}}_m^* \equiv \boldsymbol{D}_m^{-\frac{1}{2}}\boldsymbol{K}_m^*\boldsymbol{D}_m^{-\frac{1}{2}}, \tag{37}$$

$$\widetilde{\boldsymbol{K}}_d^* \equiv \boldsymbol{D}_d^{-\frac{1}{2}}\boldsymbol{K}_d^*\boldsymbol{D}_d^{-\frac{1}{2}}, \tag{38}$$

and obtain the general form of Eq. (35) and Eq. (36) as

$$\boldsymbol{L} \equiv b\boldsymbol{I} - \widetilde{\boldsymbol{K}}, \tag{39}$$

where $\widetilde{\boldsymbol{K}}$ is defined as

$$\widetilde{\boldsymbol{K}} \equiv \begin{cases} \widetilde{\boldsymbol{K}}_d^* \otimes \widetilde{\boldsymbol{K}}_m^*, \text{using Kronecker product}, \\ \widetilde{\boldsymbol{K}}_d^* \oplus \widetilde{\boldsymbol{K}}_m^*, \text{using Kronecker sum}, \end{cases} \tag{40}$$

and for the value of $b$, when the Kronecker product is used, the value of $b$ is 1; when the Kronecker sum is used, the value of $b$ is 3.

Minimizing the objective function with respect to $\widehat{\boldsymbol{Y}}$ and plugging Eq. (39) into the equation we obtain the system of linear equations as

$$vec\left(\widehat{\boldsymbol{Y}}\right) = \left((1 + b\sigma)I - \sigma\widetilde{\boldsymbol{K}}\right)^{-1} vec\left(\boldsymbol{Y}_{train}\right). \tag{41}$$

Reducing the size of the matrix without losing accuracy is helpful for saving calculation costs. For this purpose, the low-rank approximation of matrices is carried out. The approximation technique we use in this paper is based on singular value decomposition. Other matrix approximation techniques, such as incomplete Cholesky decomposition, can be employed as well. By singular value decomposition of the matrix $\boldsymbol{K}_m^*$, we can obtain

$$\boldsymbol{K}_m^* = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{T}. \tag{42}$$

According to the Eckart-Young-Mirsky theorem [62], after singular value decomposition of kernel matrix $\boldsymbol{K}_m^*$, the minimum $p - \overline{p}$ singular values of matrix $\boldsymbol{\Sigma}$ are zeroed to obtain matrix $\boldsymbol{\Sigma}_{\overline{p}}$. The parameter $\overline{p}$ is the approximate rank set by the users. The low-rank approximation of kernel matrix $\boldsymbol{K}_m^*$ can be calculated as

$$\boldsymbol{K}_m^* \approx \boldsymbol{U}_{\overline{p}}\boldsymbol{\Sigma}_{\overline{p}}\boldsymbol{V}_{\overline{p}}^{T}. \tag{43}$$

Since the kernel matrix $\boldsymbol{K}_m^*$is a symmetric positive definite matrix, we define the following matrix:

$$\boldsymbol{G}_m \equiv \boldsymbol{U}_{\overline{p}}\boldsymbol{\Sigma}_{\overline{p}}^{\frac{1}{2}}, \tag{44}$$

we can obtain

$$\boldsymbol{K}_m^* \approx \boldsymbol{U}_{\overline{p}}\boldsymbol{\Sigma}_{\overline{p}}^{\frac{1}{2}}\left(\boldsymbol{U}_{\overline{p}}\boldsymbol{\Sigma}_{\overline{p}}^{\frac{1}{2}}\right)^T \approx \boldsymbol{G}_m\boldsymbol{G}_m^T. \tag{45}$$

Similarly, the low-rank approximation of the kernel matrix $\boldsymbol{K}_d^*$ can be calculated as

$$\boldsymbol{K}_d^* \approx \boldsymbol{U}_{\overline{q}}\boldsymbol{\Sigma}_{\overline{q}}^{\frac{1}{2}}\left(\boldsymbol{U}_{\overline{q}}\boldsymbol{\Sigma}_{\overline{q}}^{\frac{1}{2}}\right)^T \approx \boldsymbol{G}_d\boldsymbol{G}_d^T. \tag{46}$$

Therefore, we can define the matrices as follows:

$$\widetilde{\boldsymbol{G}}_m \equiv \boldsymbol{D}_m^{-\frac{1}{2}}\boldsymbol{G}_m, \tag{47}$$

$$\widetilde{\boldsymbol{G}}_d \equiv \boldsymbol{D}_d^{-\frac{1}{2}}\boldsymbol{G}_d. \tag{48}$$

The normalized kernel matrices $\widetilde{\boldsymbol{K}}_m^*$ and $\widetilde{\boldsymbol{K}}_d^*$ can be written as

$$\widetilde{\boldsymbol{K}}_m^* \approx \boldsymbol{D}_m^{-\frac{1}{2}}\boldsymbol{G}_m\boldsymbol{G}_m^T\boldsymbol{D}_m^{-\frac{1}{2}} = \widetilde{\boldsymbol{G}}_m\widetilde{\boldsymbol{G}}_m^T, \tag{49}$$

$$\widetilde{\boldsymbol{K}}_d^* \approx \boldsymbol{D}_d^{-\frac{1}{2}}\boldsymbol{G}_d\boldsymbol{G}_d^T\boldsymbol{D}_d^{-\frac{1}{2}} = \widetilde{\boldsymbol{G}}_d\widetilde{\boldsymbol{G}}_d^T. \tag{50}$$

Since $\widetilde{\boldsymbol{G}}_m^T\widetilde{\boldsymbol{G}}_m \in \boldsymbol{R}^{\overline{p}\times\overline{p}}$, we can easily perform the eigendecomposition of $\widetilde{\boldsymbol{G}}_m^T\widetilde{\boldsymbol{G}}_m$ to obtain the eigenvectors and eigenvalues satisfying

$$\widetilde{\boldsymbol{G}}_m^T\widetilde{\boldsymbol{G}}_m = \overline{\boldsymbol{U}}_m diag\left(\overline{\lambda}_m^{(1)}, \overline{\lambda}_m^{(2)}, \ldots, \overline{\lambda}_m^{(\overline{p})}\right)\overline{\boldsymbol{U}}_m^T. \tag{51}$$

The eigenvectors of approximate kernel matrix $\widetilde{\boldsymbol{K}}_m^*$ may now be calculated as

$$\overline{\boldsymbol{V}}_m \equiv \widetilde{\boldsymbol{G}}_m\overline{\boldsymbol{U}}_m diag\left(\overline{\lambda}_m^{(1)}, \overline{\lambda}_m^{(2)}, \ldots, \overline{\lambda}_m^{(\overline{p})}\right)^{-\frac{1}{2}}. \tag{52}$$

Similarly, through eigendecomposition of $\widetilde{\boldsymbol{G}}_d^T\widetilde{\boldsymbol{G}}_d$, the eigenvector of approximate kernel matrix $\widetilde{\boldsymbol{K}}_d^*$ can be obtained as

$$\overline{\boldsymbol{V}}_d \equiv \widetilde{\boldsymbol{G}}_d\overline{\boldsymbol{U}}_d diag\left(\overline{\lambda}_d^{(1)}, \overline{\lambda}_d^{(2)}, \ldots, \overline{\lambda}_d^{(\overline{q})}\right)^{-\frac{1}{2}}. \tag{53}$$

Note that we have the following theorem for the eigendecomposition of the Kronecker product and Kronecker sum: if the eigenvalues of the matrices $\widetilde{\boldsymbol{K}}_m^*$ and $\widetilde{\boldsymbol{K}}_d^*$ are $\left\{\overline{\lambda}_m^{(i)}\right\}_{i=1}^{\overline{p}}$ and $\left\{\overline{\lambda}_d^{(j)}\right\}_{j=1}^{\overline{q}}$ and the corresponding eigenvectors are $\overline{\boldsymbol{V}}_m$ and $\overline{\boldsymbol{V}}_d$, then the eigenvectors of $\widetilde{\boldsymbol{K}}_m^* \otimes \widetilde{\boldsymbol{K}}_d^*$ and $\widetilde{\boldsymbol{K}}_m^* \oplus \widetilde{\boldsymbol{K}}_d^*$ are both

$$\overline{\boldsymbol{V}} \equiv \overline{\boldsymbol{V}}_d \otimes \overline{\boldsymbol{V}}_m, \tag{54}$$

and their eigenvalues are as follows:

$$\left[\overline{\wedge}\right]_{i,j} \equiv \begin{cases} \overline{\lambda}_m^{(i)}\overline{\lambda}_d^{(j)}, \text{using Kronecker product,} \\ \overline{\lambda}_m^{(i)} + \overline{\lambda}_d^{(j)}, \text{using Kronecker sum.} \end{cases} \tag{55}$$

We note that

$$\overline{\boldsymbol{V}}^T\overline{\boldsymbol{V}} = \boldsymbol{I}, \tag{56}$$

and by using the Woodbury Equation [63], we can compute the target function as

$$vec\left(\widehat{\boldsymbol{Y}}\right) = \frac{1}{1+b\sigma}vec\left(\boldsymbol{Y}_{train}\right) + \frac{1}{(1+b\sigma)^2}\overline{\boldsymbol{V}}diag\left(vec\left(\overline{\boldsymbol{D}}\right)\right)\overline{\boldsymbol{V}}^T vec\left(\boldsymbol{Y}_{train}\right), \tag{57}$$

where $\overline{\boldsymbol{D}}$ is defined as

$$\left[\overline{D}\right]_{i,j} \equiv \left(\frac{1}{\sigma\left[\overline{\wedge}\right]_{i,j}} - \frac{1}{1+b\sigma}\right)^{-1} = \frac{\sigma(1+b\sigma)\left[\overline{\wedge}\right]_{i,j}}{1+b\sigma-\sigma\left[\overline{\wedge}\right]_{i,j}}. \tag{58}$$

Then, we can get rid of the vec operator, and using vec-trick techniques [60,64], obtained the final state of the solution as

$$\widehat{\boldsymbol{Y}} = \frac{1}{1+b\sigma}\boldsymbol{Y}_{train} + \frac{1}{(1+b\sigma)^2}\overline{\boldsymbol{V}}_m\left(\overline{\boldsymbol{D}} * \left(\overline{\boldsymbol{V}}_m^T\boldsymbol{Y}_{train}\overline{\boldsymbol{V}}_d\right)\right)\overline{\boldsymbol{V}}_d^T, \tag{59}$$

where * is the Hadamard product operator.

In summary, we obtained two types of Link Propagation: Link Propagation using the Kronecker sum (LP-S) and Link Propagation using the Kronecker product (LP-P).

The overview of the approach we propose is summarized in Algorithm 1.

## 3 Results

### 3.1 Evaluation measurements

Two assessment criteria, area under the curve (AUC) and area under the precision-recall curve (AUPR), were used in our study to assess the model performance. These criteria are also widely used in related studies [28,65−67]. The area under the receiver operating characteristic (ROC) curve is calculated by graphing the true positive rate versus the false-positive rate at various threshold values to produce the AUC. The AUC ranges from 0 to 1 but is rarely less than 0.5. Generally, if the AUC value is greater than 0.5, we believe that the performance is better than random guessing. A high AUC value indicates good model performance. The area under the curve formed by graphing precision against recall at various threshold levels is known as the AUPR. As with the AUC, AUPR should also be greater than 0 and less than 1, and the higher its value, the better the model works.

### 3.2 Parameter selection

Grid search machine learning is one of the most common, intuitive, and effective parameter adjustment methods and is used to obtain the optimal parameter under 5-CV. Link propagation contains three parameters that need to be adjusted,

**Algorithm 1** Algorithm of HSIC-MKL+LP

**Require:** The masked matrix $\boldsymbol{Y}_{train} \in \boldsymbol{R}^{p\times q}$, the miRNA similarity kernels $\boldsymbol{K}_{m,1}$, $\boldsymbol{K}_{m,2}$, $\boldsymbol{K}_{m,3} \in \boldsymbol{R}^{p\times p}$ and the disease similarity kernels $\boldsymbol{K}_{d,1}$, $\boldsymbol{K}_{d,2}$, $\boldsymbol{K}_{d,3} \in \boldsymbol{R}^{q\times q}$;Three parameters $\sigma$, $\overline{p}$, and $\overline{q}$ for link propagation;

**Ensure:** The prediction of link strength matrix $\widehat{\boldsymbol{Y}} \in \boldsymbol{R}^{p\times q}$;

1: Computing the miRNA and disease similarity kernels, which are list in Table 1;

2: Computing the miRNA kernel weights $\boldsymbol{\eta}_m$ and disease kernel weights $\boldsymbol{\eta}_d$ by using HSIC-MKL, respectively, and using the weights to fuse the kernels, and obtained the optimal miRNA kernel $\boldsymbol{K}'_m$ and the optimal disease kernel $\boldsymbol{K}'_d$;

3: Using Eq. (28) to normalize the optimal kernels, and get the normalized miRNA kernel $\boldsymbol{K}^*_m$ and the normalized disease kernel $\boldsymbol{K}^*_d$;

4: Computing the low-rank approximation of the matrix $\boldsymbol{K}^*_m$ and $\boldsymbol{K}^*_d$ by Eq. (44) and Eq. (45);

5: Computing the matrices $\widetilde{\boldsymbol{G}}_m$ and $\widetilde{\boldsymbol{G}}_d$ by Eq. (46) and Eq. (47);

6: Computing the eigen decomposition of $\widetilde{\boldsymbol{G}}_m^T\widetilde{\boldsymbol{G}}_m$ and $\widetilde{\boldsymbol{G}}_d^T\widetilde{\boldsymbol{G}}_d$ as in Eq. (50), and get the eigenvalues and the corresponding eigenvectors;

7: Using Eq. (57) to compute the matrix $\overline{\boldsymbol{D}}$;

8: Estimating the $\widehat{\boldsymbol{Y}}$ by Eq. (58).

which are $\sigma$, $\overline{p}$ and $\overline{q}$. $\sigma$ is the regularization parameter and should have a value greater than 0. The $\overline{p}$ and $\overline{q}$ represent the rank of the approximate matrix of the miRNA kernel and disease kernel, respectively. Hence, $\overline{p}$ and $\overline{q}$ should both be positive integers, less than 495 and 383, respectively. In grid search, the variation trend of AUPR value and AUC value of model LP-S and LP-P with the values of parameters $\overline{p}$ and $\overline{q}$ was represented by heat map, as shown in Fig. 4. As mentioned above, Link Propagation involves two types of using the Kronecker product and the Kronecker sum. Therefore, we have two different sets of parameters. When using the LP-S, the optimal combination of parameters is $\sigma = 0.2$, $\overline{p} = 175$, and $\overline{q} = 325$. In addition, $\sigma = 0.25$, $\overline{p} = 175$, and $\overline{q} = 325$ is for LP-P.

### 3.3 Performance analysis

To test the validity of the model, we established LP-S and LP-P models with HSIC-MKL on the MDA dataset under 5-CV. In order to compare the performance of HSIC-MKL and other multi-core learning methods, we choose Centered Kernel Alignment-based Multiple Kernel Learning (CKA-MKL) algorithm, which also measures the relationship between kernel matrices. CKA-MKL also has theoretical support and has been well performed in previous studies [68−71]. Additionally, a mean weight fusion technique was applied to show that HSIC-MKL can fuse kernels more effectively. The results of our approach are shown in Table 2, and the comparison of AUC and AUPR of the four models is shown in Fig. 5.

The results suggest that the model using LP-S with HSIC-MKL achieves the best results, and its AUC and AUPR are 0.9800 and 0.8374, respectively. In addition, the models using the Kronecker sum perform better than models using the Kronecker product. The results also show that the CKA-MKL performance lags behind that of HSIC-MKL, while the mean weight performs the worst, which also proves the effectiveness of HSIC-MKL. And the reason for that is HSIC-MKL performs better than CKA-MKL in combining data from different sources and measuring the relationship between the target matrix and the kernel matrices.

In addition, the weights of miRNA and disease kernels calculated by HSIC-MKL are shown in Fig. 6. The $\boldsymbol{K}_{m,2}$, which contains functional information of miRNA, has the largest weight, followed by $\boldsymbol{K}_{m,3}$. The weight of $\boldsymbol{K}_{m,1}$ is close to zero, which means that the sequence information of miRNA

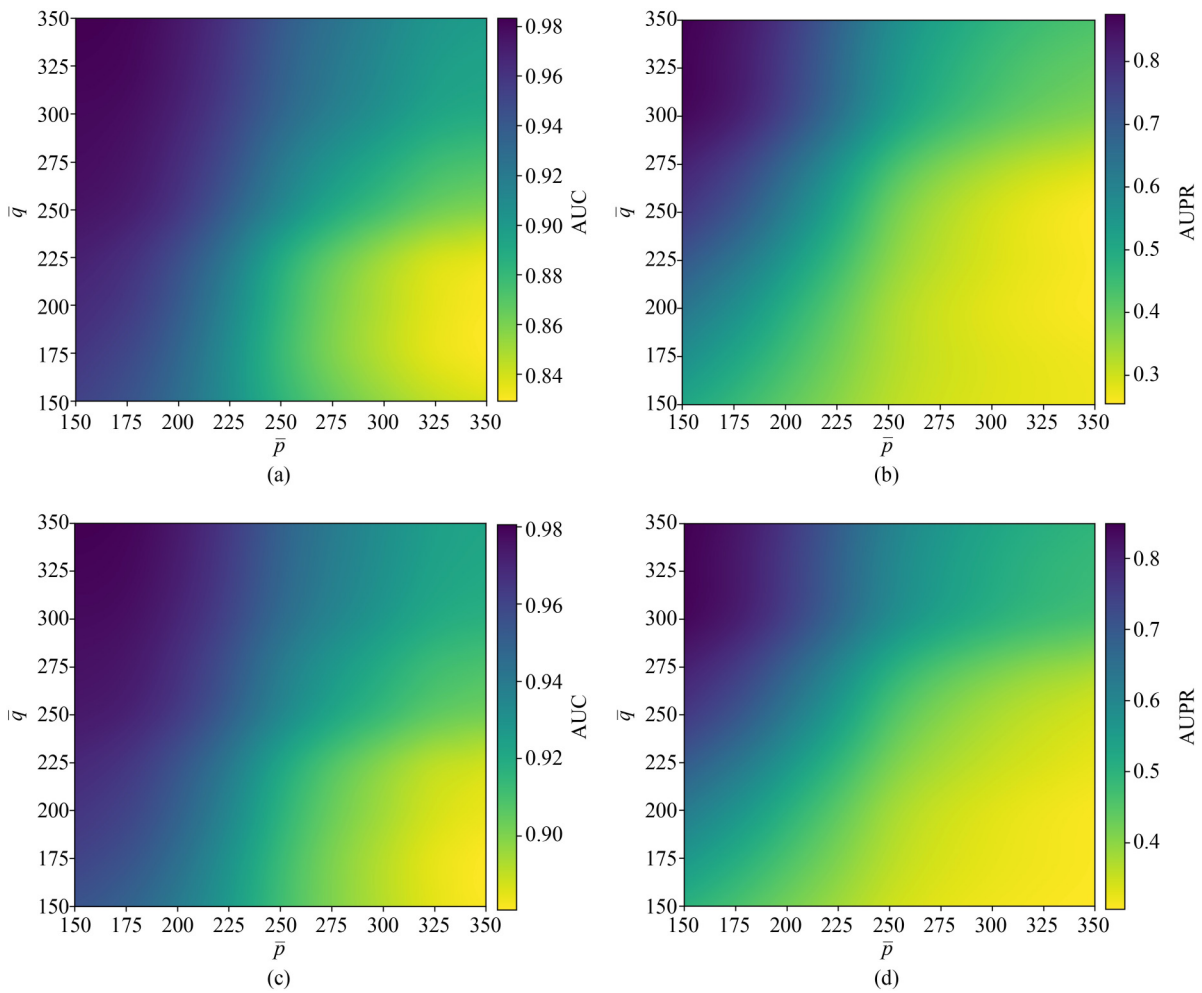



**Fig. 4** The AUC and AUPR of different models under 5-CV. (a) The AUC of parameter $\overline{p}$ and $\overline{q}$ of HSIC-MKL + LP-S; (b) the AUPR of parameter $\overline{p}$ and $\overline{q}$ of HSIC-MKL + LP-S; (c) the AUC of parameter $\overline{q}$ and $\overline{q}$ of HSIC-MKL + LP-P; (d) the AUPR of parameter $\overline{p}$ and $\overline{q}$ of HSIC-MKL + LP-P

**Table 2** The performance (AUC and AUPR) of different models under 5-CV

| Model | AUC | AUPR |
|---|---|---|
| HSIC-MKL + LP-S | **0.9800** | **0.8374** |
| HSIC-MKL + LP-P | 0.9781 | 0.8144 |
| CKA-MKL + LP-S | 0.9790 | 0.8136 |
| CKA-MKL + LP-P | 0.9768 | 0.7963 |
| Mean weighted + LP-S | 0.9761 | 0.8074 |
| Mean weighted + LP-P | 0.9738 | 0.7849 |

hardly plays a role. Each of the three kernel weights for disease has a value close to one-third. The weights of $\boldsymbol{K}_{d,2}$ and $\boldsymbol{K}_{d,3}$ are close, followed by $\boldsymbol{K}_{d,1}$. The greater the weight calculated by the multiple kernel learning method, the greater the contribution of this kernel the prediction of MDA, and the greater attention should be paid to this information in biological experiments.

### 3.4 Comparison with other methods

To assess the performance of HSIC-MKL+LP-S and highlight the effectiveness of our work, we compared our approach to the following approaches: CKA-HGTMF [72], CKA-MKL+HGBLM [28], FKL-Spa-LapRLS [65], MDA-SKF [56], LRSSLMDA [73], PBMDA [74], MCMDA [75], NCPMDA [37], RLSMDA [76], HDMP [77], and WBSMDA [26]. Our methods and the approaches mentioned above are listed in Table 3. According to the results, HSIC-MKL+LP-S performs best under 5-CV, obtains the highest AUC and AUPR values and surpasses existing methods.

### 3.5 Local CV and global CV

When a disease is newly discovered, there is no known miRNA information associated with it, so we can assess the effectiveness of predicting miRNA for new diseases using local LOOCV. In addition, the performance of the model is

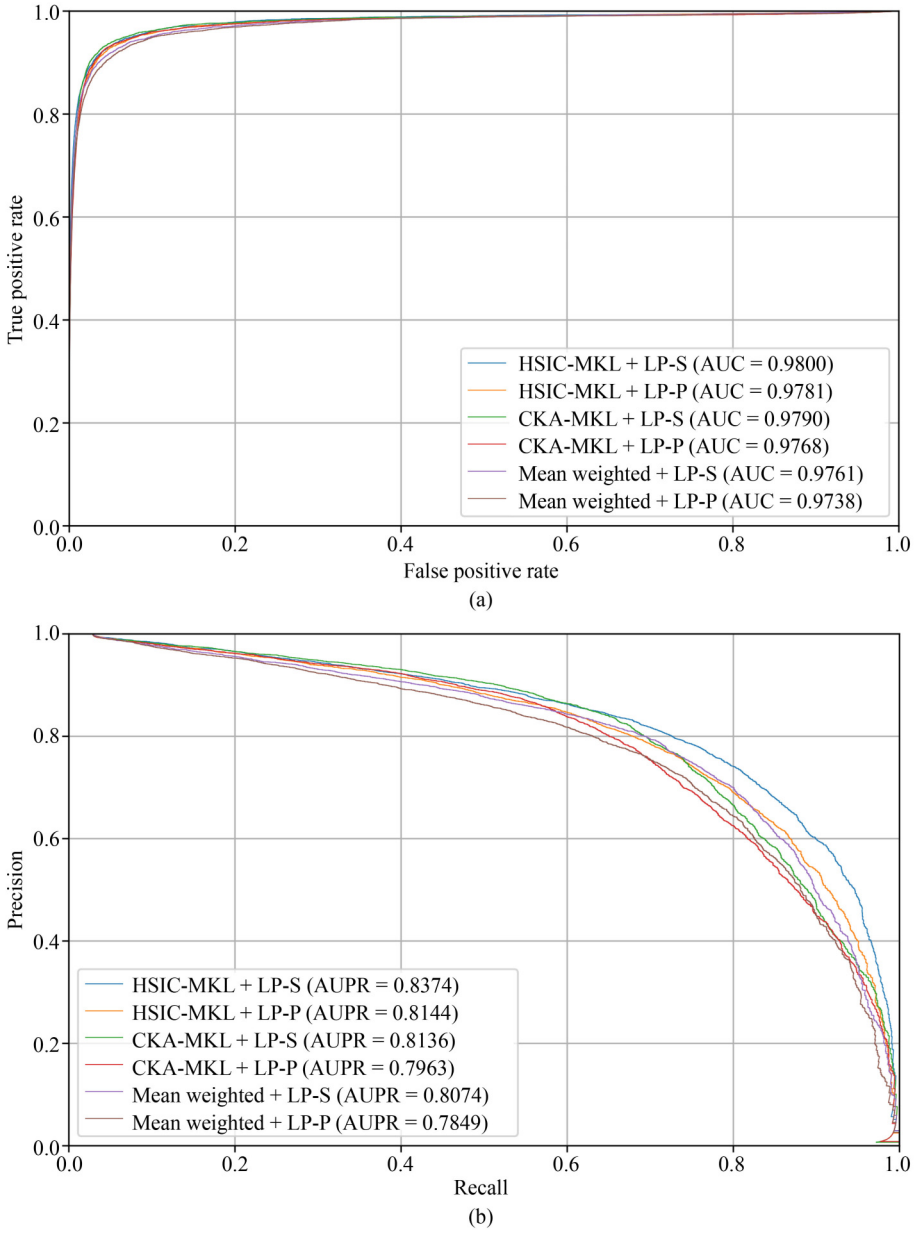



**Fig. 5** The AUC and AUPR of different models under 5-CV. (a) AUC; (b) AUPR

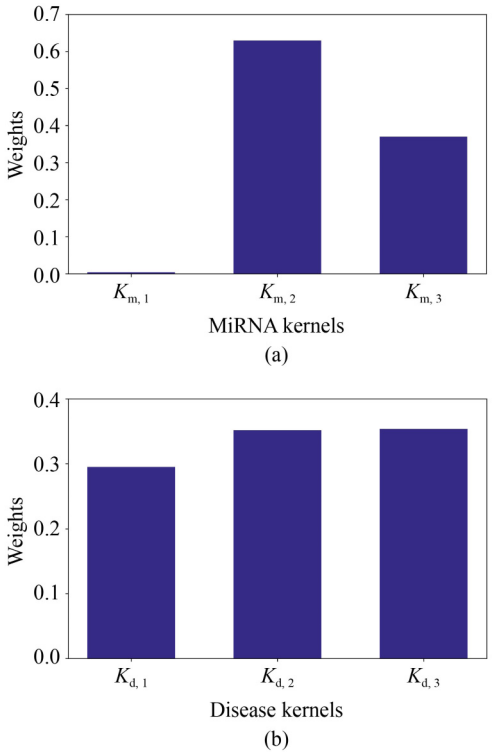



**Fig. 6** The kernel weights of miRNA and disease. (a) MiRNA kernel weights; (b) disease kernel weights

further assessed using global LOOCV. Generally, local LOOCV and global LOOCV are evaluated only by AUC, so the AUC of our results and existing methods are listed in Table 4. CKA-MKL+HGBLM [28], FKL-Spa-LapRLS [65], MDA-SKF [56], LRSSLMDA [73], PBMDA [74], MCMDA [75], NCPMDA [37], RLSMDA [76], HDMP [77], and WBSMDA [26] are also included for comparison

We found that HSIC-MKL+LP-S is not the best local LOOCV. In order to save the cost of calculations and enhance the overall effectiveness of the model, our approach uses matrix approximation, which leads to poor miRNA prediction results for diseases with few miRNA associations, thus lowering the overall effect. However, of the models we compared ours performs better than some. In addition, our results show that HSIC-MKL+LP-S has the best performance in global LOOCV, as its AUC value reaches 0.9874.

### 3.6 Running time

The computational efficiency of the algorithm is also an important index to evaluate the algorithm, so we analyze the time complexity of the algorithm and test the running time of different models through experiments. The time complexity of both the LP-S and LP-P methods we use in this paper is $O(\overline{p}^3+\overline{q}^3)$. As we can see, the running time of the algorithm depends on the approximate rank set by the users, which means that the user can significantly reduce the computation time by adjusting the parameters.

In addition, we separately test the running time of multiple kernel learning methods and LP under 5-CV. On a computer with AMD Ryzen 5 3600 (3.6 GHz 6-core 12-threads CPU), 16GB memory and Windows OS, we run ten experiments, and the results are averaged. The results of running time are shown in Table 5. Among the three multi-kernel learning methods, mean weighted method is very simple, its running time is close to 0, and HSIC-MKL is much faster than CKA-MKL. The running times of algorithm LP-S and LP-P are roughly the same, and algorithm LP-S is a little faster.

**Table 3** Comparison on different methods under 5-CV

| Model | AUC | AUPR |
|---|---|---|
| HSIC-MKL+LP-S (our method) | **0.9800** | **0.8374** |
| CKA-HGRTMF | 0.9775 | 0.7712 |
| CKA-MKL+HGBLM | 0.9680 | 0.7291 |
| MDA-SKF | 0.9557 | – |
| FKL-Spa-LapRLS | 0.9535 | – |
| LRSSLMDA | 0.9181 | – |
| PBMDA | 0.9172 | – |
| MCMDA | 0.8767 | – |
| NCPMDA | 0.8763 | – |
| RLSMDA | 0.8569 | – |
| HDMP | 0.8342 | – |
| WBSMDA | 0.8185 | – |

The best results in each column are in bold faces.

**Table 4** Comparison on different methods via Global LOOCV and Local LOOCV

| Model | Global LOOCV | Local LOOCV |
|---|---|---|
| HSIC-MKL + LP-S (our method) | **0.9874** | 0.7800 |
| CKA-MKL + HGBLM | 0.9715 | 0.8083 |
| MDA-SKF | 0.9576 | 0.8356 |
| FKL-Spa-LapRLS | 0.9563 | **0.8398** |
| LRSSLMDA | 0.9178 | 0.8418 |
| PBMDA | 0.9169 | 0.8341 |
| MCMDA | 0.8749 | 0.7718 |
| NCPMDA | 0.9073 | 0.8584 |
| RLSMDA | 0.8426 | 0.6953 |
| HDMP | 0.8366 | 0.7702 |
| WBSMDA | 0.8030 | 0.8031 |

The best results in each column are in bold faces.

## 4 Discussion and conclusion

In this paper, we propose a link propagation model using matrix decomposition and low-rank approximation. Our model was inspired by label propagation. The kernels in miRNA and disease space are fused using HSIC-MKL.

Our method is compared with existing methods, including CKA-HGTMF, CKA-MKL+HGBLM, FKL-Spa-LapRLS, MDA-SKF, LRSSLMDA, PBMDA, MCMDA, NCPMDA, RLSMDA, HDMP, and WBSMDA. The results show that under 5-CV, our model exceeds the AUPR and AUC values of all existing methods, and the effect is significantly improved. In local LOOCV, our model is weaker than several existing models but it also achieves the best results in global LOOCV.

There are many existing methods for predicting MDA,

**Table 5** The running time (seconds) of different models under 5-CV

| Model | Running time/s |
|---|---|
| HSIC-MKL | 11.50 |
| CKA-MKL | 21.33 |
| Mean weighted | 0.07 |
| LP-S | 7.56 |
| LP-P | 7.88 |

however, a link propagation algorithm based on matrix decomposition and matrix approximation technology improves the performance of the prediction of MDA. With this approach, we can guarantee model accuracy while also saving on the cost of calculations. In addition, if a suitable similarity kernel is constructed, our method can also be used in other bioinformatics domains, including the identification of drug-side effect associations. Of course, our model has many limitations, such as its performance in local LOOCV, which is less accurate than some models. Further work is still needed to improve the model. Chen et al. [78] summarize the current issues and future trends. Dataset is an important problem that limits the development of computational models. The current dataset contains 495 miRNAs and 383 diseases, but the known associations is only 2.86%. At the same time, negative MDA is not available. Furthermore, the development of computational models should focus on miRNAs that do not have any known associated diseases, as new miRNAs are frequently discovered in biology. In addition, we should pay more attention to deep learning. Graph convolutional networks is operated on graph-based data, which can further improve the performance of the prediction of MDA.

Biologists have access to publicly available source codes and datasets at the website of github.com/Kinkou626/HSIC-MKL-LP.

**Acknowledgements** This work was supported in part by the National Natural Science Foundation of China (Grant Nos. 62072385, 62172076, and U22A2038), the Municipal Government of Quzhou (2022D040), and the Zhejiang Provincial Natural Science Foundation of China (No. LY23F020003).

## References

1. Shi H, Zhang G, Zhou M, Cheng L, Yang H, Wang J, Sun J, Wang Z. Integration of multiple genomic and phenotype data to infer novel miRNA-disease associations. PLoS One, 2016, 11(2): e0148521
2. Carthew R W, Sontheimer E J. Origins and mechanisms of miRNAs and siRNAs. Cell, 2009, 136(4): 642–655
3. Peng Y, Liu Y, Chen X. Bioinformatics analysis reveals functions of MicroRNAs in rice under the drought stress. Current Bioinformatics, 2020, 15(8): 927–936
4. Roehle A, Hoefig K P, Repsilber D, Thorns C, Ziepert M, Wesche K O, Thiere M, Loeffler M, Klapper W, Pfreundschuh M, Matolcsy A, Bernd H W, Reiniger L, Merz H, Feller A C. MicroRNA signatures characterize diffuse large B‐cell lymphomas and follicular lymphomas. British Journal of Haematology, 2008, 142(5): 732–744
5. Cogswell J P, Ward J, Taylor I A, Waters M, Shi Y, Cannon B, Kelnar K, Kemppainen J, Brown D, Chen C, Prinjha R K, Richardson J C, Saunders A M, Roses A D, Richards C A. Identification of miRNA changes in Alzheimer's disease brain and CSF yields putative biomarkers and insights into disease pathways. Journal of Alzheimer's Disease, 2008, 14(1): 27–41
6. Caporali A, Meloni M, Völlenkle C, Bonci D, Sala-Newby G B, Addis R, Spinetti G, Losa S, Masson R, Baker A H, Agami R, Le Sage C, Condorelli G, Madeddu P, Martelli F, Emanueli C. Deregulation of microRNA-503 contributes to diabetes mellitus−induced impairment of endothelial function and reparative angiogenesis after limb ischemia. Circulation, 2011, 123(3): 282–291
7. Hu Y, Zhang Y, Zhang H, Gao S, Wang L, Wang T, Han Z, Sun B, Liu G. Cognitive performance protects against Alzheimer’s disease independently of educational attainment and intelligence. Molecular Psychiatry, 2022, 27(10): 4297−4306
8. Anonymous. 2021 Alzheimer’s disease facts and figures. Alzheimer’s & Dement, 2021, 17(3): 327−406
9. Hu Y, Sun J, Zhang Y, Zhang H, Gao S, Wang T, Han Z, Wang L, Sun B L, Liu G. rs1990622 variant associates with Alzheimer's disease and regulates *TMEM106B* expression in human brain tissues. BMC Medicine, 2021, 19(1): 11
10. Hu Y, Zhang H, Liu B, Gao S, Wang T, Han Z, International Genomics of Alzheimer’s Project (IGAP), Ji X, Liu G. rs34331204 regulates *TSPAN13* expression and contributes to Alzheimer's disease with sex differences. Brain, 2020, 143(11): e95
11. Bhaumik D, Scott G K, Schokrpur S, Patil C K, Campisi J, Benz C C. Expression of microRNA-146 suppresses NF-κB activity with reduction of metastatic potential in breast cancer cells. Oncogene, 2008, 27(42): 5643–5647
12. Wang N, Li Y, Liu S, Gao L, Liu C, Bao X, Xue P. Analysis and validation of differentially expressed MicroRNAs with their target genes involved in GLP-1RA facilitated osteogenesis. Current Bioinformatics, 2021, 16(7): 928–942
13. Hu Y, Qiu S, Cheng L. Integration of multiple-Omics data to analyze the population-specific differences for coronary artery disease. Computational and Mathematical Methods in Medicine, 2021, 2021: 7036592
14. Hu Y, Zhang Y, Zhang H, Gao S, Wang L, Wang T, Han Z, International Genomics of Alzheimer's Project (IGAP), Liu G. Mendelian randomization highlights causal association between genetically increased C-reactive protein levels and reduced Alzheimer's disease risk. Alzheimer’s & Dement, 2022, 18(10): 2003–2006
15. Tang W, Wan S, Yang Z, Teschendorff A E, Zou Q. Tumor origin detection with tissue-specific miRNA and DNA methylation markers. Bioinformatics, 2018, 34(3): 398–406
16. Sarkar J P, Saha I, Sarkar A, Maulik U. Machine learning integrated ensemble of feature selection methods followed by survival analysis for predicting breast cancer subtype specific miRNA biomarkers. Computers in Biology and Medicine, 2021, 131: 104244
17. Zhu Q, Fan Y, Pan X. Fusing multiple biological networks to effectively predict miRNA-disease associations. Current Bioinformatics, 2021, 16(3): 371–384
18. Zhang Y, Duan G, Yan C, Yi H, Wu F X, Wang J. MDAPlatform: a component-based platform for constructing and assessing miRNA-disease association prediction methods. Current Bioinformatics, 2021, 16(5): 710–721
19. Chen X, Zhu C C, Yin J. Ensemble of decision tree reveals potential miRNA-disease associations. PLoS Computational Biology, 2019, 15(7): e1007209
20. Fu H, Huang F, Liu X, Qiu Y, Zhang W. MVGCN: data integration through multi-view graph convolutional network for predicting links in biomedical bipartite networks. Bioinformatics, 2022, 38(2): 426–434
21. Zhang G, Li M, Deng H, Xu X, Liu X, Zhang W. SGNNMD: signed graph neural network for predicting deregulation types of miRNA-disease associations. Briefings in Bioinformatics, 2022, 23(1): bbab464
22. Huang F, Yue X, Xiong Z, Yu Z, Liu S, Zhang W. Tensor decomposition with relational constraints for predicting multiple types of microRNA-disease associations. Briefings in Bioinformatics, 2021, 22(3): bbaa140
23. Lu X, Gao Y, Zhu Z, Ding L, Wang X, Liu F, Li J. A constrained probabilistic matrix decomposition method for predicting miRNA-disease associations. Current Bioinformatics, 2021, 16(4): 524–533
24. Lan W, Dong Y, Chen Q, Liu J, Wang J, Chen Y P P, Pan S. IGNSCDA: predicting CircRNA-disease associations based on improved graph convolutional network and negative sampling. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2022, 19(6): 3530–3538
25. Peng W, Che Z, Dai W, Wei S, Lan W. Predicting miRNA-disease associations from miRNA-gene-disease heterogeneous network with multi-relational graph convolutional network model. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2022, doi:

10.1109/TCBB.2022.3187739
26. Chen X, Yan C, Zhang X, You Z H, Deng L, Liu Y, Zhang Y, Dai Q. WBSMDA: within and between score for MiRNA-disease association prediction. Scientific Reports, 2016, 6(1): 21106
27. Chen X, Yin J, Qu J, Huang L. MDHGI: matrix decomposition and heterogeneous graph inference for miRNA-disease association prediction. PLoS Computational Biology, 2018, 14(8): e1006418
28. Ding Y, Jiang L, Tang J, Guo F. Identification of human microRNA-disease association via hypergraph embedded bipartite local model. Computational Biology and Chemistry, 2020, 89: 107369
29. Chen X, Wang L, Qu J, Guan N N, Li J Q. Predicting miRNA−disease association based on inductive matrix completion. Bioinformatics, 2018, 34(24): 4256–4265
30. Chen X, Sun L G, Zhao Y. NCMCMDA: miRNA−disease association prediction through neighborhood constraint matrix completion. Briefings in Bioinformatics, 2021, 22(1): 485–496
31. Fu L, Peng Q. A deep ensemble model to predict miRNA-disease association. Scientific Reports, 2017, 7(1): 14482
32. Zeng X, Ding N, Rodríguez-Patón A, Zou Q. Probability-based collaborative filtering model for predicting gene−disease associations. BMC Medical Genomics, 2017, 10(S5): 76
33. Zeng X, Liu L, Lü L Y, Zou Q. Prediction of potential disease-associated microRNAs using structural perturbation method. Bioinformatics, 2018, 34(14): 2425–2432
34. Zeng X, Wang W, Deng G, Bing J, Zou Q. Prediction of potential disease-associated MicroRNAs by using neural networks. Molecular Therapy Nucleic Acids, 2019, 16: 566–575
35. Chen X, Liu M X, Yan G Y. RWRMDA: predicting novel human microRNA−disease associations. Molecular BioSystems, 2012, 8(10): 2792–2798
36. Van Laarhoven T, Nabuurs S B, Marchiori E. Gaussian interaction profile kernels for predicting drug−target interaction. Bioinformatics, 2011, 27(21): 3036–3043
37. Gu C, Liao B, Li X, Li K. Network consistency projection for human miRNA-disease associations inference. Scientific Reports, 2016, 6: 36054
38. Tiwari P, Dehdashti S, Obeid A K, Marttinen P, Bruza P. Kernel method based on non-linear coherent states in quantum feature space. Journal of Physics A: Mathematical and Theoretical, 2022, 55(35): 355301
39. Li Y, Qiu C, Tu J, Geng B, Yang J, Jiang T, Cui Q. HMDD v2. 0: a database for experimentally supported human microRNA and disease associations. Nucleic Acids Research, 2014, 42(D1): D1070–D1074
40. Chen X, Li T H, Zhao Y, Wang C C, Zhu C C. Deep-belief network for predicting potential miRNA-disease associations. Briefings in Bioinformatics, 2021, 22(3): bbaa186
41. Wang C C, Li T H, Huang L, Chen X. Prediction of potential miRNA−disease associations based on stacked autoencoder. Briefings in Bioinformatics, 2022, 23(2): bbac021
42. Kozomara A, Griffiths-Jones S. miRBase: annotating high confidence microRNAs using deep sequencing data. Nucleic Acids Research, 2014, 42(D1): D68–D73
43. 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(13): 1644–1650
44. Zhu C C, Wang C C, Zhao Y, Zuo M, Chen X. Identification of miRNA−disease associations via multiple information integration with Bayesian ranking. Briefings in Bioinformatics, 2021, 22(6): bbab302
45. Zhao Y, Chen X, Yin J. Adaptive boosting-based computational model for predicting potential miRNA-disease associations. Bioinformatics, 2019, 35(22): 4730–4738
46. Lowe H J, Barnett G O. Understanding and using the medical subject headings (MeSH) vocabulary to perform literature searches. JAMA, 1994, 271(14): 1103–1108
47. Luo J, Xiao Q, Liang C, Ding P. Predicting MicroRNA-disease associations using Kronecker regularized least squares based on heterogeneous omics data. IEEE Access, 2017, 5: 2503–2513
48. Lan W, Wang J, Li M, Liu J, Wu F X, Pan Y. Predicting microRNA-disease associations based on improved microRNA and disease similarities. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2018, 15(6): 1774–1782
49. Lee I, Blom U M, Wang P I, Shim J E, Marcotte E M. Prioritizing candidate disease genes by network-based boosting of genome-wide association data. Genome Research, 2011, 21(7): 1109–1121
50. Cheng L, Wang G, Li J, Zhang T, Xu P, Wang Y. SIDD: a semantically integrated database towards a global view of human disease. PLoS One, 2013, 8(10): e75504
51. Gretton A, Bousquet O, Smola A, Schölkopf B. Measuring statistical dependence with Hilbert-Schmidt norms. In: Proceedings of the 16th International Conference on Algorithmic Learning Theory. 2005: 63–77
52. Wang T, Li W. Kernel learning and optimization with Hilbert−Schmidt independence criterion. International Journal of Machine Learning and Cybernetics, 2018, 9(10): 1707–1717
53. Xuan J, Lu J, Yan Z, Zhang G. Bayesian deep reinforcement learning via deep kernel learning. International Journal of Computational Intelligence Systems, 2018, 12(1): 164–171
54. Wang T, Lu J, Zhang G. Two-stage fuzzy multiple kernel learning based on Hilbert−Schmidt independence criterion. IEEE Transactions on Fuzzy Systems, 2018, 26(6): 3703–3714
55. Gönen M, Alpaydın E. Multiple kernel learning algorithms. The Journal of Machine Learning Research, 2011, 12: 2211–2268
56. Jiang L, Ding Y, Tang J, Guo F. MDA-SKF: similarity kernel fusion for accurately discovering miRNA-disease association. Frontiers in Genetics, 2018, 9: 618
57. Zhou D, Bousquet O, Lal T N, Weston J, Schölkopf B. Learning with local and global consistency. In: Proceedings of the 16th International Conference on Neural Information Processing Systems. 2003: 321–328
58. Zhu X, Ghahramani Z, Lafferty J. Semi-supervised learning using Gaussian fields and harmonic functions. In: Proceedings of the 20th International Conference on Machine Learning (ICML-03). 2003: 912–919
59. Raymond R, Kashima H. Fast and scalable algorithms for semi-supervised link prediction on static and dynamic graphs. In: Proceedings of 2010 European Conference on Machine Learning and Knowledge Discovery in Databases. 2010: 131–147
60. Laub A J. Matrix Analysis for Scientists and Engineers. Philadelphia: SIAM, 2005
61. Kashima H, Kato T, Yamanishi Y, Sugiyama M, Tsuda K. Link propagation: a fast semi-supervised learning algorithm for link prediction. In: Proceedings of the 9th SIAM International Conference on Data Mining. 2009: 1093–1104
62. Golub G H, Hoffman A, Stewart G W. A generalization of the Eckart-Young-Mirsky matrix approximation theorem. Linear Algebra and its Applications, 1987, 88–89: 317-327
63. Bishop C M, Nasrabadi N M. Pattern Recognition and Machine Learning. New York: Springer, 2006
64. Vishwanathan S V N, Borgwardt K M, Schraudolph N N. Fast computation of graph kernels. In: Proceedings of the 19th International Conference on Neural Information Processing Systems. 2006
65. Jiang L, Xiao Y, Ding Y, Tang J, Guo F. FKL-Spa-LapRLS: an accurate method for identifying human microRNA-disease association. BMC Genomics, 2018, 19(S10): 911
66. Ding Y, Tiwari P, Zou Q, Guo F, Pandey H M. C-loss based higher order fuzzy inference systems for identifying DNA N4-methylcytosine sites. IEEE Transactions on Fuzzy Systems, 2022, 30(11): 4754–4765
67. Chen X, Xie D, Wang L, Zhao Q, You Z H, Liu H. BNPMDA: bipartite network projection for MiRNA−disease association prediction. Bioinformatics, 2018, 34(18): 3178–3186
68. Cristianini N, Shawe-Taylor J, Elisseeff A, Kandola J. On kernel-target alignment. In: Proceedings of the 14th International Conference on Neural Information Processing Systems. 2001: 367–373

69. Cortes C, Mohri M, Rostamizadeh A. Algorithms for learning kernels based on centered alignment. The Journal of Machine Learning Research, 2012, 13(1): 795–828
70. Lu Y, Wang L, Lu J, Yang J, Shen C. Multiple kernel clustering based on centered kernel alignment. Pattern Recognition, 2014, 47(11): 3656–3664
71. Hu J, Li Y, Zhang M, Yang X, Shen H B, Yu D J. Predicting protein-DNA binding residues by weightedly combining sequence-based features and boosting multiple SVMs. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2017, 14(6): 1389–1398
72. Wang H, Tang J, Ding Y, Guo F. Exploring associations of non-coding RNAs in human diseases via three-matrix factorization with hypergraph-regular terms on center kernel alignment. Briefings in Bioinformatics, 2021, 22(5): bbaa409
73. Chen X, Huang L. LRSSLMDA: laplacian regularized sparse subspace learning for MiRNA-disease association prediction. PLoS Computational Biology, 2017, 13(12): e1005912
74. You Z H, Huang Z A, Zhu Z, Yan G Y, Li Z W, Wen Z, Chen X. PBMDA: a novel and effective path-based computational model for miRNA-disease association prediction. PLoS Computational Biology, 2017, 13(3): e1005455
75. Li J Q, Rong Z H, Chen X, Yan G Y, You Z H. MCMDA: matrix completion for MiRNA-disease association prediction. Oncotarget, 2017, 8(13): 21187–21199
76. Chen X, Yan G Y. Semi-supervised learning for potential human microRNA-disease associations inference. Scientific Reports, 2014, 4: 5501
77. Xuan P, Han K, Guo M, Guo Y, Li J, Ding J, Liu Y, Dai Q, Li J, Teng Z, Huang Y. Prediction of microRNAs associated with human diseases based on weighted *k* most similar neighbors. PLoS One, 2013, 8(8): e70204
78. Chen X, Xie D, Zhao Q, You Z H. MicroRNAs and complex diseases: from experimental results to computational models. Briefings in Bioinformatics, 2019, 20(2): 515–539

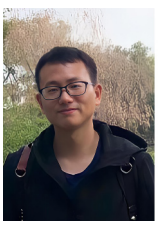
Yizheng Wang is a postgraduate in the Institute of Fundamental and Frontier Sciences, University of Electronic Science and Technology of China, China. He received the Bachelor of Engineering degree in computer science and technology from Yanshan University, China in 2022. His research interests include bioinformatics and machine learning.

Xin Zhang is a deputy chief physician of Beidahuang Industry Group General Hospital, China. He graduated from Harbin Medical University, China in 2006, and his research direction is basic medicine and lung cancer.

Ying Ju received her PhD degree in Biomedical Engineering from Xi'an Jiaotong University, China. She is an associate professor with the Department of Computer Science, Xiamen University, China. She has published more than 15 papers in journal and conference. Her main research interest is biomedical engineering.

Qing Liu is a chief physician of Department of Anesthesiology, Hospital (T.C.M) Affiliated to Southwest Medical University, China. He received his master's degree in Medicine in 2004, and his research interest is mechanism of neuropathic pain.

Quan Zou received the BSc, MSc, and the PhD degrees in computer science from Harbin Institute of Technology, China in 2004, 2007 and 2009, respectively. He is currently a professor in the Institute of Fundamental and Frontier Sciences, University of Electronic Science and Technology of China. His research is in the areas of bioinformatics, machine learning and parallel computing. Several related works have been published by Science, Briefings in Bioinformatics, Bioinformatics, etc. Google scholar showed that his more than 100 papers have been cited more than 16000 times. He is the editor-in-chief of Current Bioinformatics and Computers in Biology and Medicine. He was selected as one of the Clarivate Analytics Highly Cited Researchers in 2018–2022.



Yazhou Zhang received PhD degree in computer applications technology from Tianjin University, China in 2020. He has published more than 35 papers, including CCF ranking A/B conference papers (e.g., IJCAI, EMNLP, CIKM, NAACL) and top journal papers (e.g., IEEE Trans. on Fuzzy System, Information Fusion, ACM Trans. on Internet Technology, Theoretical Computer Science, Neural Networks).

Yijie Ding received the PhD degree in computer science from the School of Computer Science and Technology, Tianjin University, China in 2018. He is currently an Associate Professor with the Yangtze Delta Region Institute, University of Electronic Science and Technology of China, China. His research interests include bioinformatics and machine learning. Several related works have been published by Briefings in Bioinformatics, IEEE TFS, IEEE TAI, IEEE/ACM TCBB, IEEE JBHI, Information Sciences, Knowledge-Based Systems, Applied Soft Computing, and Neurocomputing.

Ying Zhang is a chief physician of Department of Anesthesiology, Hospital (T.C.M) Affiliated to Southwest Medical University, China. She is studying for her PhD at Macau University of Science and Technology, China. She received her master's degree in Medicine in 2011, and her research interest is mechanism of neuropathic pain and protective mechanism of postoperative cognitive function.