

RESEARCH

Open Access



# SPLHRNMTF: robust orthogonal non-negative matrix tri-factorization with self-paced learning and dual hypergraph regularization for predicting miRNA-disease associations

Dong Ouyang<sup>1\*</sup>, Rui Miao<sup>2</sup>, Juan Zeng<sup>1</sup>, Xing Li<sup>1</sup>, Ning Ai<sup>3</sup>, Panke Wang<sup>1</sup>, Jie Hou<sup>1</sup> and Jinqiu Zheng<sup>1</sup>

## Abstract

MicroRNAs (miRNAs) have been demonstrated to be closely related to human diseases. Studying the potential associations between miRNAs and diseases contributes to our understanding of disease pathogenic mechanisms. As traditional biological experiments are costly and time-consuming, computational models can be considered as effective complementary tools. In this study, we propose a novel model of robust orthogonal non-negative matrix tri-factorization (NMTF) with self-paced learning and dual hypergraph regularization, named SPLHRNMTF, to predict miRNA-disease associations. More specifically, SPLHRNMTF first uses a non-linear fusion method to obtain miRNA and disease comprehensive similarity. Subsequently, the improved miRNA-disease association matrix is reformulated based on weighted  $k$ -nearest neighbor profiles to correct false-negative associations. In addition, we utilize  $L_{2,1}$  norm to replace Frobenius norm to calculate residual error, alleviating the impact of noise and outliers on prediction performance. Then, we integrate self-paced learning into NMTF to alleviate the model from falling into bad local optimal solutions by gradually including samples from easy to complex. Finally, hypergraph regularization is introduced to capture high-order complex relations from hypergraphs related to miRNAs and diseases. In 5-fold cross-validation five times experiments, SPLHRNMTF obtains higher average AUC values than other baseline models. Moreover, the case studies on breast neoplasms and lung neoplasms further demonstrate the accuracy of SPLHRNMTF. Meanwhile, the potential associations discovered are of biological significance.

**Keywords** miRNA-disease associations, Non-negative matrix tri-factorization, Self-paced learning, Hypergraph regularization,  $L_{2,1}$  norm

## Introduction

MicroRNAs (miRNAs) are a class of non-coding RNA molecules (about 22 nucleotides in length) encoded by endogenous genes [1]. MiRNAs are involved in post-transcriptional gene expression regulation and have been demonstrated to play an important role in many human life processes, including cell proliferation [2], cell growth [3], tumor invasion [4], immune response [5], and so on. In addition, more and more studies have shown that

\*Correspondence:

Dong Ouyang  
ouyangdong@gdmu.edu.cn

<sup>1</sup> School of Biomedical Engineering, Guangdong Medical University, Dongguan 523808, China

<sup>2</sup> Basic Teaching Department, Zhuhai Campus of Zunyi Medical University, Zhuhai 519099, China

<sup>3</sup> The college of Mechanical and Electrical Engineering, Shihezi University, Shihezi 832003, China



© The Author(s) 2024. **Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

miRNA dysregulation often leads to the occurrence of various human diseases. For example, miR-145 inhibits the cell proliferation of human lung adenocarcinoma by targeting EGFR and NUDT1 [6]. The levels of miR-105 in blood and tumors are related to ZO-1 expression and metastatic progression in breast cancer [7]. The expression of miR-34a is excessive in specific brain regions of Alzheimer's disease patients as well as in the 3xTg-AD mouse model [8]. Therefore, identifying disease-related miRNAs contributes to revealing the pathogenesis of diseases and discovering potential biomarkers at the miRNA level. Since the identification of miRNA-disease associations using traditional wet experiments is time-consuming and expensive, more and more computational methods have been proposed as efficient complementary tools to infer potential associations between miRNAs and diseases.

Over the past few decades, many computational methods have been proposed based on the assumption that miRNAs with similar functions are associated with phenotypically similar diseases. Chen et al. [9] proposed the RWRMDA method, which predicts potential miRNA-disease associations by performing random walk with restart on the miRNA-miRNA functional similarity network. However, this method is not suitable for miRNAs with no association, which limits its performance improvement. Xuan et al. [10] designed a new prediction method called HDMP, which combines the distribution of disease-related miRNAs in  $k$  neighborhoods with miRNA function similarity to predict potential associations between miRNAs and diseases. However, HDMP only considers local network similarity, resulting in sub-optimal prediction performance. Xuan et al. [11] further developed a new prediction method to infer miRNA-disease associations based on random walk on miRNA functional similarity network. Chen et al. [12] designed the within and between score for miRNA-disease association prediction (WBSMDA) method to identify disease-related miRNAs by integrating various biological similarity networks. You et al. [13] developed a path-based PBMDA method and used a depth-first search algorithm to predict miRNA-disease associations. Chen et al. [14] presented a bipartite heterogeneous network link prediction method based on co-neighbor to infer the associations between miRNAs and diseases. Li et al. [15] developed the NPRWR method to predict miRNA-disease associations by using dual random walk with restart and network projection technology. Chen et al. [16] presented a deep belief network model called DBNMDA based on Restricted Boltzmann Machines for miRNA-disease association prediction. Ha et al. [17] proposed a novel computational model of metric learning named MLMD for inferring miRNA-disease associations. Based

on deep neural networks, Ha et al. [18] further designed a new node2vec-based neural collaborative filtering framework named NCMD to predict the associations between miRNAs and diseases. Based on graph convolutional networks, Li et al. [19] proposed a neural inductive matrix completion method, called NIMCGCN, to infer miRNA-disease associations. Tang et al. [20] also employed graph convolutional networks and multichannel attention mechanism to extract and enhance latent representations of miRNAs and diseases. They predicted potential associations based on the reconstructed miRNA-disease association matrix. Recently, Peng et al. [21] used hypergraph convolutional networks and a variational auto-encoder to develop a MHCLMDA model with contrastive learning for predicting the associations between miRNAs and diseases.

In recent years, many studies have shown that matrix factorization methods with feature extraction and reconstruction capabilities are also considered as promising computational methods. For example, Xiao et al. [22] proposed a non-negative matrix factorization method with graph Laplacian regularization constraint, named GRNMF, to predict potential miRNA-disease associations. Chen et al. [23] designed an IMCMDA method of inductive matrix completion, which constrains integrated miRNA similarity and disease similarity into matrix factorization to predict the associations between miRNAs and diseases. Chen et al. [24] proposed a novel matrix decomposition and heterogeneous graph inference (MDHGI) method for identifying miRNA-disease associations. MDHGI integrates the association probability derived from matrix decomposition through sparse learning method and biological similarity information into a heterogeneous network. Dissez et al. [25] developed a drug repositioning method based on non-negative matrix tri-factorization (NMTF) with graph Laplacian regularization, which utilizes NMTF to simultaneously decompose multiple binary association matrices between indications and drugs, drugs and proteins, proteins and pathways, and drugs and diseases. Jamali et al. [26] presented an NMTF-DTI method of non-negative matrix tri-factorization with multiple kernel fusion for drug-target interaction prediction. This method makes use of graph Laplacian regularization to preserve the local geometric structure of the biological similarity network in low-dimensional space. Zheng et al. [27] proposed the NMFMC method based on non-negative matrix factorization, which decomposes the association matrix into a known part and an unknown part for miRNA-disease association prediction. Ha et al. [28] designed a new method based on probabilistic matrix factorization (PMF), called IMIPME, to identify miRNA-disease associations. Subsequently, Ha [29] also developed a new

computational framework (MDMF) using matrix factorization with disease similarity constraints to identify potential miRNA-disease associations. Through further considering miRNA and disease similarities and integrating them into matrix factorization, Ha [30] designed a simple and effective computational framework, called SMAP, to predict miRNA-disease associations. Recently, Ha [31] also proposed a novel matrix decomposition method named EMFLDA, which applies lncRNA expression profiles to identify the associations between lncRNAs and diseases.

Although the above matrix decomposition-based methods have achieved excellent performance in potential association prediction, most methods still have some limitations. First, it is challenging to reasonably integrate multi-view similarity of miRNAs (or diseases) into a comprehensive similarity network. Most methods employ a linear strategy based on one similarity to integrate the similarity of different views. More specifically, multiple similarities are integrated by using one similarity to fill the missing parts of another similarity or by averaging different types of similarity. However, simple linear fusion can not effectively learn more complex integrated information, thus limiting the improvement of prediction performance. Second, limited by the current level of science and technology, miRNA-disease association data usually have “association, but we don’t know yet” (i.e., false-negative samples) associations in a large number of negative samples, which significantly affects the model prediction performance. Third, most matrix decomposition-based methods utilize Frobenius norm to calculate residual error, which may cause the prediction performance to be affected by noise and outliers. Fourth, because methods based on matrix decomposition usually have non-convexity, these methods can easily obtain a bad local optimal solution. Finally, the pairwise relations within homogeneous graphs are usually considered by most methods while ignoring high-order complex relations present in heterogeneous graphs.

To alleviate the limitations mentioned above, we propose a novel computational model of **Self-Paced Learning and Hypergraph regularization into Robust orthogonal Non-negative Matrix Tri-Factorization (SPLHRNMTF)** to identify the associations between miRNAs and diseases, which preserves double orthogonality conditions and dual hypergraph regularization. Compared with simple linear fusion, we first propose a non-linear fusion method to integrate different types of miRNA or disease similarity network into a comprehensive similarity network for miRNAs or diseases. Then, the weighted  $k$ -nearest neighbor profiles are used to obtain an improved miRNA-disease association matrix, thereby effectively correcting false-negative samples. Next, we separately

introduce  $L_{2,1}$  norm and hypergraph regularization into NMTF to effectively alleviate the influence of noise and outliers on prediction performance, and capture the high-order complex relations from hypergraphs related to miRNAs and diseases. Finally, in order to cope with the fact that NMTF with non-convexity easily obtains a bad local solution, we integrate self-paced learning into NMTF to effectively alleviate the problem of falling into such solution by gradually including samples from easy to complex. To solve the optimization problem of SPLHRNMTF, we propose an alternating optimization algorithm, whose convergence is theoretically guaranteed. The 5-fold cross-validation is performed five times to evaluate the proposed SPLHRNMTF model. Specifically, SPLHRNMTF can obtain better prediction performance than other matrix factorization-based models, even superior to some graph and hypergraph convolutional network-based models. Furthermore, the results of ablation experiment indicate that each module in the SPLHRNMTF and preprocessing step are effective. In addition, we conduct case studies on breast neoplasms and lung neoplasms, 48 and 46 of top 50 are confirmed by experimental reports. Moreover, unconfirmed potential associations have biological significance. In summary, SPLHRNMTF model demonstrates efficacy and accuracy in predicting associations between miRNAs and diseases.

## Materials

### Human miRNA-disease association datasets

The Human MiRNA Disease Database (HMDD) is a database containing experimentally verified human miRNA-disease associations [32, 33]. In this work, HMDD v2.0 and HMDD v3.2 were used as benchmark databases to construct association matrices, where these two databases can be downloaded from <http://www.cuilab.cn/hmdd>. To more extensively investigate the generalization performance of the model, we also obtained a dataset named MDAv3.2\_data1 based on the study of Ning et al. [34]. In addition, we further preprocessed HMDD v2.0 and HMDD v3.2 databases to obtain MDAv2.0\_data and MDAv3.2\_data2 datasets, respectively. To be more specific, we removed these miRNAs that had no sequence in the miRBase database [35] and were not found in the MISIM 2.0 database [36]. Meanwhile, we deleted these diseases that were not found and whose category was not “C” in Medical Subject Headings (MeSH) tree structure. Finally, the first dataset MDAv2.0\_data contains 5896 experimentally verified human miRNA-disease associations between 548 miRNAs and 320 diseases. The second dataset MDAv3.2\_data1 involves 853 miRNAs, 591 diseases and 12446 experimentally verified associations. The third dataset MDAv3.2\_data2 includes 9676

experimentally confirmed associations between 812 miRNAs and 469 diseases.

**Problem formulation**

Given a set of miRNAs  $\mathcal{M} = \{m_1, m_2, \dots, m_m\}$  and a set of diseases  $\mathcal{D} = \{d_1, d_2, \dots, d_n\}$ , we can construct the miRNA-disease associations into a matrix  $X \in \{0, 1\}^{m \times n}$ , where  $m$  and  $n$  represent the size of the set of miRNAs and diseases, respectively. Obviously, an entry  $x_{ij}$  of the matrix  $X$  is set to 1 if a miRNA is associated with a disease. Otherwise, the unknown or unobserved entries are set to 0. In this work, we transformed the identification of miRNA-disease associations into a matrix completion task. However, the matrix  $X$  is extremely sparse with a large number of unknown and unobserved entries, which is a challenge to efficiently perform matrix completion. Therefore, miRNA similarity information and disease similarity information were considered as auxiliary information to alleviate the challenge.

**Disease semantic similarity**

The disease semantic similarity was calculated by using directed acyclic graph (DAG) with disease hierarchical relationships obtained from the MeSH database (<https://www.nlm.nih.gov/mesh/>). The DAG of a disease  $d_i$  can be defined as  $DAG(d_i) = (d_i, T(d_i), E(d_i))$ , where  $T(d_i)$  denotes the set of  $d_i$  and its ancestor nodes,  $E(d_i)$  represents the set of edges with regard to the direct links between parent nodes and child nodes. Then, the first semantic contribution of diseases  $d_k$  to  $d_i$  can be defined as follows:

$$SC^1(d_i, d_k) = \begin{cases} 1, & \text{if } d_k = d_i \\ \max\{\Delta \times SC^1(d_i, d'_k) | d'_k \in \text{children of } d_k\}, & \text{if } d_k \neq d_i. \end{cases} \tag{1}$$

where  $\Delta$  is a semantic contribution decay factor, and we set it to 0.5 according to previous work [37]. Then, the semantic value of disease  $d_i$  can be calculated as below:

$$SV^1(d_i) = \sum_{d_k \in T(d_i)} SC^1(d_i, d_k) \tag{2}$$

On the basis of the assumption that if the intersection of a disease pair in the DAG is larger, their semantic similarity value will be greater. Then, the first disease semantic similarity  $SD^1(d_i, d_j)$  between diseases  $d_i$  and  $d_j$  can be defined as follows:

$$SD^1(d_i, d_j) = \frac{\sum_{d_t \in T(d_i) \cap T(d_j)} (SC^1(d_i, d_t) + SC^1(d_j, d_t))}{SV^1(d_i) + SV^1(d_j)} \tag{3}$$

However,  $SD^1$  fails to consider the significance of semantic contributions of different diseases. It overlooks

the fact that diseases appearing in fewer DAG may be more specific, and thus should have higher semantic contribution values. According to previous work [23], the second semantic contribution of disease  $d_k$  to  $d_i$  can be presented as below:

$$SC^2(d_i, d_k) = -\log \left( \frac{\text{the number of DAGs including } d_k}{\text{the number of disease}} \right) \tag{4}$$

Similarly, we can define the second semantic value  $SV^2(d_i)$  of disease  $d_i$  and the second disease semantic similarity  $SD^2(d_i, d_j)$  between diseases  $d_i$  and  $d_j$  as follows:

$$SV^2(d_i) = \sum_{d_k \in T(d_i)} SC^2(d_i, d_k) \tag{5}$$

$$SD^2(d_i, d_j) = \frac{\sum_{d_t \in T(d_i) \cap T(d_j)} (SC^2(d_i, d_t) + SC^2(d_j, d_t))}{SV^2(d_i) + SV^2(d_j)} \tag{6}$$

Based on previous study [38], we integrated these two kinds of disease semantic similarity  $SD^1(d_i, d_j)$  and  $SD^2(d_i, d_j)$  to obtain a more reasonable disease semantic similarity. Finally, the final disease semantic similarity  $SD_1(d_i, d_j)$  between diseases  $d_i$  and  $d_j$  can be presented as below:

$$SD_1(d_i, d_j) = \frac{SD^1(d_i, d_j) + SD^2(d_i, d_j)}{2} \tag{7}$$

**MiRNA functional similarity**

On the basis of the assumption that functionally similar miRNAs tend to be associated with phenotypically similar diseases and vice versa, we can calculate miRNA functional similarity scores [37]. In this study, we obtained miRNA functional similarity from the MISIM 2.0 database (<http://www.lirmed.com/misim/>). Then, we constructed a matrix  $SM_1$  by using these data, where  $SM_1(m_i, m_j)$  denotes the functional similarity score between miRNA  $m_i$  and  $m_j$ .

**MiRNA sequence similarity**

According to previous work [39], we obtained the miRNA sequences containing ‘‘AUCG’’ from the miR-Base database (<https://www.mirbase.org/>). Based on the sequence information, we utilized pairwise sequence alignment function ‘‘pairwiseAlignment’’ in R package *Biostrings* to calculate miRNA sequence similarity score. To be more specific, the gap opening penalty is set to 5, the gap extension penalty is set to 2, the match score is set to 1 and the mismatch score is set to -1 in this function. Finally, the miRNA sequence similarity matrix  $SM_2$  can be obtained by *min-max* normalization as follows:

$$SM_2(m_i, m_j) = \frac{Score(m_i, m_j) - Score_{min}}{Score_{max} - Score_{min}} \quad (8)$$

where  $Score_{min}$  and  $Score_{max}$  denote the minimum and maximum similarity score of all miRNA pairs, respectively.  $Score(m_i, m_j)$  indicates that sequence similarity score can be calculated by using function “pairwiseAlignment”.

### Gaussian interaction profile kernel similarity for miRNAs and diseases

On the basis of assumption that miRNAs with similar functions are more likely to be associated with similar diseases, we can calculate the Gaussian interaction profile kernel similarity through the known miRNA-disease association network [40]. For a given miRNA  $m_i$ , its interaction profile  $IP(m_i)$  was extracted from the known miRNA-disease association matrix. Then, the Gaussian interaction profile kernel similarity for miRNAs  $SM_3(m_i, m_j)$  between miRNAs  $m_i$  and  $m_j$  can be presented as follows:

$$SM_3(m_i, m_j) = \exp(-\gamma_m \|IP(m_i) - IP(m_j)\|^2) \quad (9)$$

where  $\gamma_m$  controls the kernel bandwidth, which can be calculated by using the following equation:

$$\gamma_m = \frac{\gamma'_m}{\frac{1}{m} \sum_{i=1}^m \|IP(m_i)\|^2} \quad (10)$$

where  $m$  represents the number of miRNAs and  $\gamma'_m$  is set to 1 based on previous study [40]. Similarly, the Gaussian interaction profile kernel similarity for diseases  $SD_2(d_i, d_j)$  between diseases  $d_i$  and  $d_j$  can be obtained based on the following two equations:

$$SD_2(d_i, d_j) = \exp(-\gamma_d \|IP(d_i) - IP(d_j)\|^2) \quad (11)$$

$$\gamma_d = \frac{\gamma'_d}{\frac{1}{n} \sum_{i=1}^n \|IP(d_i)\|^2} \quad (12)$$

where  $IP(d_i)$  is a binary vector extracted from the known miRNA-disease association matrix,  $n$  represents the number of diseases and  $\gamma'_d$  is also set to 1.

### Comprehensive similarity for miRNAs and diseases

Based on previous works [12, 41], integrating different similarity information can not only avoid being too one-sided, but also obtain more accurate and reasonable integrated similarity for miRNAs and diseases. However, the simple linear similarity combination method is challenging to apply for the fusion of multiple biological similarity

information. Inspired by previous study [42], we utilized a non-linear fusion method to integrate multiple similarity networks into a single similarity network for miRNAs and diseases, respectively. Next, we took the integration of multiple miRNA similarity information as an example to introduce the construction of comprehensive miRNA similarity.

In the first step, we performed a better normalization to calculate normalized weight matrix  $SM'_\nu$  for the  $\nu$ -th type of similarity network as follows:

$$SM'_\nu(m_i, m_j) = \begin{cases} \frac{SM_\nu(m_i, m_j)}{2 \sum_{k \neq i} SM_\nu(m_i, m_k)}, & j \neq i \\ \frac{1}{2}, & j = i \end{cases} \quad (13)$$

In the second step, we employed  $k$ -nearest neighbor (KNN) to measure the local relationship of each similarity network. The detailed calculation process was as below:

$$K_\nu(m_i, m_j) = \begin{cases} \frac{SM_\nu(m_i, m_j)}{\sum_{k \in N_i} SM_\nu(m_i, m_k)}, & j \in N_i \\ 0, & otherwise \end{cases} \quad (14)$$

where  $N_i$  is a set of  $k$  nearest neighbors for the node  $m_i$  of miRNA similarity networks.  $K_\nu$  denotes the local affinity kernel of the  $\nu$ -th data type. In this paper, according to previous study [42], we set the neighbor parameter in KNN method approximately as  $|N|/10$ , where  $|N|$  is the total number of nodes in the similarity network.

In the third step, we iteratively updated the similarity matrix for each type of data using the following procedure:

$$SM'^{(t+1)}_\nu = K_\nu \times \frac{\sum_{k \neq \nu} SM'^{(t)}_k}{M - 1} \times (K_\nu)^T \quad (15)$$

where  $\nu = \{1, 2, \dots, M\}$  and  $M$  denotes the total number of data types.  $SM'^{(t+1)}_\nu$  is the status matrix of the  $\nu$ -th data type after  $t$  iterations.

In this work, the iteration stops when Eq. (15) reaches a convergence criterion, defined as a relative change  $\frac{\|SM'^{(t+1)}_\nu - SM'^{(t)}_\nu\|}{\|SM'^{(t)}_\nu\|}$  being less than  $10^{-5}$ . After iterative update, the ultimate similarity matrix  $SM$  could be obtained as follows:

$$SM = \frac{SM'^{(t)}_1 + SM'^{(t)}_2 + \dots + SM'^{(t)}_M}{M} \quad (16)$$

According to the above update rules, the obtained similarity matrix is not a symmetric matrix, so we further calculated  $SM = \frac{SM + SM^T}{2}$  as the miRNA comprehensive similarity matrix. Similarly, we can get the disease comprehensive similarity matrix  $SD$  based on the same calculation rules of miRNA.

### Methods

In this paper, we proposed the SPLHRNMTF model that integrates self-paced learning and hypergraph regularization into NMTF using  $L_{2,1}$  norm for predicting the associations between miRNAs and diseases. Figure 1 illustrates the whole workflow of SPLHRNMTF.

#### Self-paced learning

In recent years, self-paced learning (SPL) has emerged as a successful approach in mitigating the problem of falling into local optimal solutions, owing to its training strategy of gradually selecting samples from simple to complex [43, 44]. Given a training dataset  $\mathcal{D} = \{x_i, y_i\}_{i=1}^N$ , the objective function of SPL was as follows:

$$\min_{\theta, w} \sum_{i=1}^N w_i L(f(x_i, \theta), y_i) + g(\eta, w) \tag{17}$$

where  $L(\cdot, \cdot)$  denotes the training error calculated by the loss function,  $\theta$  is the model parameter of the loss function,  $w = [w_1, w_2, \dots, w_N]^T$  represents a weight variable, and  $\eta$  is the learning pace parameter.

To better reflect the potential importance of samples and realize the advantages of SPL, a novel SPL regularization term was utilized [45]. The regularization term  $g(\eta, w)$  was defined as follows:

$$g(\eta, w) = - \sum_{i=1}^N \zeta \ln(w_i + \zeta/\eta) \tag{18}$$

Based on previous work [45], the optimal  $w^*$  can be calculated as follows:

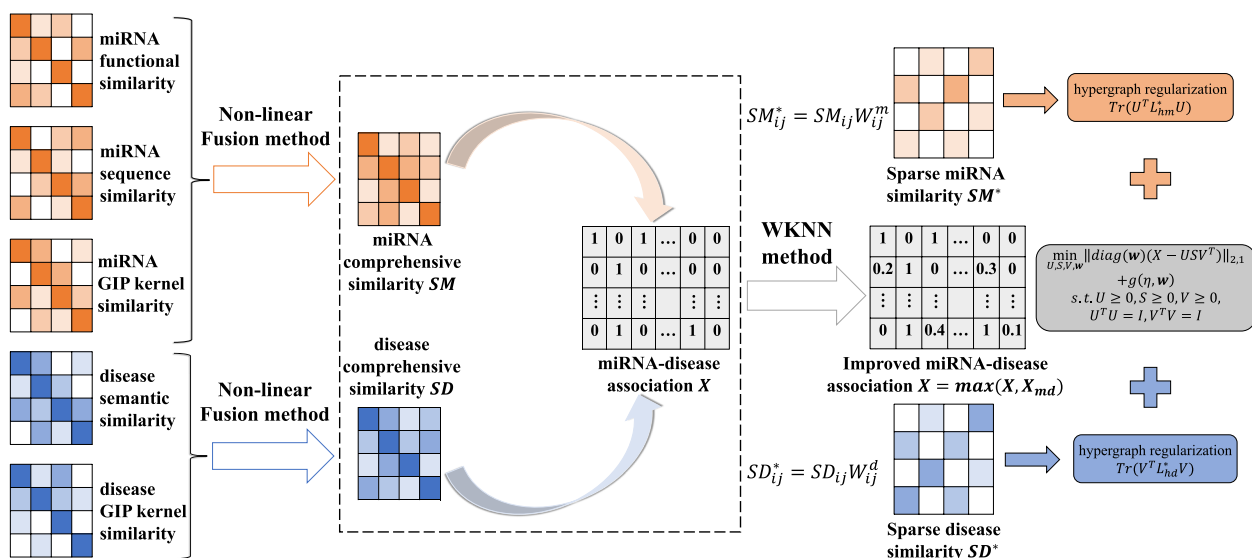
$$w_i^* = \begin{cases} 1, & \text{if } l_i \leq \zeta\eta/(\zeta + \eta) \\ 0, & \text{if } l_i \geq \eta \\ \zeta/l_i - \zeta/\eta, & \text{otherwise.} \end{cases} \tag{19}$$

where  $\zeta$  is set to  $0.5 \times \eta$  for simplicity in our experiments. With the increase of  $\eta$ , more and more samples will be selected until all samples are chosen.

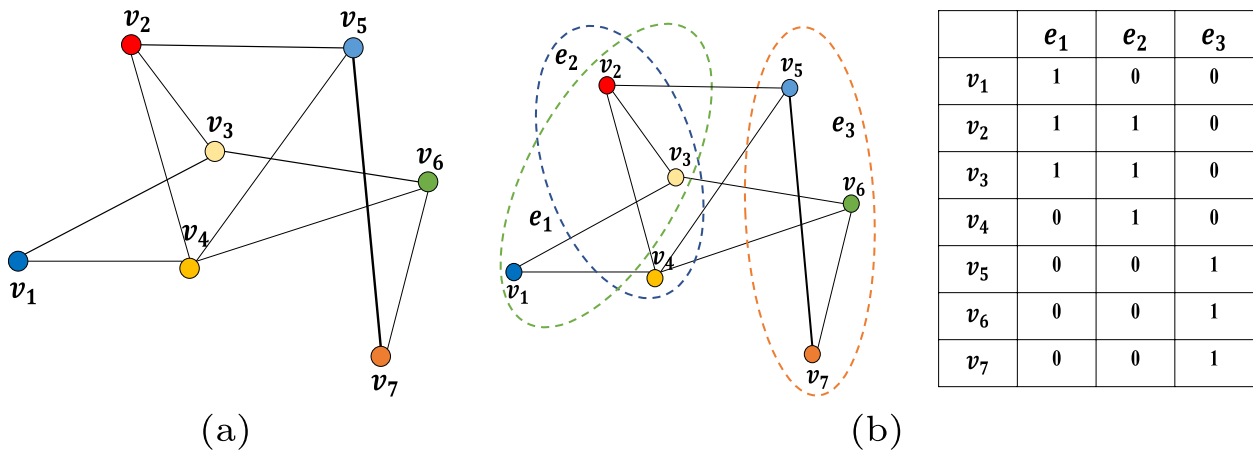
#### Hypergraph learning

Recent studies on manifold learning theory and spectral graph theory shows that when original data points were mapped from high-dimensional space to low-dimensional space,  $k$ -nearest neighbor graphs can effectively preserve the potential geometric structure of high-dimensional data in low-dimensional space [46]. However, the classic graph structure only considers the pairwise relationships of data points in the neighbor graph, while ignoring high-order relationships. To alleviate this problem, hypergraph, as a generalization of graph, has been proposed for data representation [47, 48]. Figure 2 is an example that shows the difference between classic graph and hypergraph.

Given  $G = (V, E, W)$  is a weighted hypergraph, which contains a finite hyperedge set  $E = \{e_i | i = 1, 2, \dots, \bar{m}\}$  and a finite vertex set  $V = \{v_j | j = 1, 2, \dots, \bar{n}\}$ . For the construction of hyperedge, we utilized the KNN method to learn hypergraphs for miRNAs and diseases, respectively. For example, we concatenated miRNA-disease associations and miRNA comprehensive similarity as



**Fig. 1** The workflow of our proposed SPLHRNMTF model for predicting potential miRNA-disease associations. Note that GIP kernel similarity denotes Gaussian interaction profile kernel similarity



**Fig. 2** The difference between classic graph and hypergraph. **a** A classic graph. **b** A hypergraph and its incidence matrix

features of node miRNAs. Based on the concatenated features, we calculated the nearest  $k$  neighbors of each miRNA based on Euclidean distance, thereby determining a subset (i.e., hyperedge) from the  $k$  neighbors. In addition, each hyperedge  $e$  is a subset of  $V$ .  $W$  is a diagonal matrix that indicates the weight of hyperedges. The incidence matrix  $H$  of  $G$  can be defined as below:

$$H(v, e) = \begin{cases} 1, & \text{if } v \in e \\ 0, & \text{if } v \notin e. \end{cases} \quad (20)$$

In this work, we initialized the weight value of each hyperedge in the hypergraph by constructing an affinity matrix. Specifically, the affinity matrix  $A$  of hypergraph was calculated as follows:

$$A_{ij} = \exp\left(-\frac{\|v_i - v_j\|^2}{\sigma^2}\right) \quad (21)$$

where  $\sigma$  is standard deviation distance among all vertices. Then, the initial weight of each hyperedge can be defined as follows:

$$W_i = \sum_{v_j \in e_i} A_{ij} \quad (22)$$

The degree of each vertex  $v \in V$  in incidence matrix  $H$  was computed as follows:

$$d_V(v) = \sum_{e \in E} w(e)H(v, e) \quad (23)$$

The degree of each hyperedge  $e \in E$  was defined as the number of vertices included in  $E$ . Specifically,  $d_E(e)$  can be denoted as:

$$d_E(e) = \sum_{v \in V} H(v, e) \quad (24)$$

Let  $F$  denotes the low-dimensional embedding representation of miRNAs (or diseases), we can obtain the following hypergraph structure loss function:

$$\begin{aligned} \Omega &= \frac{1}{2} \sum_{e \in E} \sum_{(v_i, v_j) \in V} \frac{w(e)H(v_i, e)H(v_j, e)}{d_E(e)} \|F(v_i) - F(v_j)\|_F^2 \\ &= \text{Tr}\left(F^T (D_V - HWD_E^{-1}H^T) F\right) \\ &= \text{Tr}\left(F^T L_{\text{hyper}} F\right) \end{aligned} \quad (25)$$

where  $D_V$  denotes a diagonal matrix where its elements related to the degree of vertices.  $D_E$  can be defined as a diagonal matrix whose elements correspond to the degree of hyperedges.  $L_{\text{hyper}} = D_V - S$ , and  $S = HWD_E^{-1}H^T$ .

### Improved miRNA-disease association

The values of interaction profiles without known association are all zeros, but they could be potential true associations (i.e., false-negative samples). This may result in an unsatisfactory prediction performance. To alleviate the above problems, we performed preprocessing step to calculate new interaction profiles by using weighted  $k$ -nearest neighbor (WKNN) profiles.

For each miRNA  $m_p$ , we conducted a sorting of all other miRNAs in descending order, considering their similarity with  $m_p$ . Subsequently, the similarity of  $m_p$  with its  $K$  nearest known miRNAs was computed, ensuring that each of miRNAs has at least one known association. Finally, its similarity information was combined with the corresponding  $K$  interaction profiles, resulting in a new interaction profile as follows:

$$X_m(m_p) = \frac{1}{E_m} \sum_{i=1}^K \theta^{i-1} \times SM(m_i, m_p) \times X(m_i) \quad (26)$$

where  $E_m = \sum_{1 \leq i \leq K} SM(m_i, m_p)$  and  $\theta \in [0, 1]$  are the normalization term and decay term, respectively.

Similar with miRNA, the new interaction profiles for each disease  $d_q$  can be calculated as below:

$$X_d(d_q) = \frac{1}{E_d} \sum_{j=1}^K \theta^{j-1} \times SD(d_j, d_q) \times X(d_j) \quad (27)$$

where  $E_d = \sum_{1 \leq j \leq K} SD(d_j, d_q)$ .

Finally, we obtained the new interaction profiles of miRNAs and diseases by using the average of  $X_m$  and  $X_d$ . The miRNA-disease adjacency matrix is updated as follows:

$$X = \max(X, X_{md}) \quad (28)$$

where  $X_{md} = (X_m + X_d)/2$ .

### SPLHRNMTF

Based on previous works [49, 50], two biological entities in the same cluster are more likely to have similar characteristics, so the sparse similarity matrix constructed based on  $k$ -nearest neighbors has been effectively applied. The weighted matrix  $W^m$  can be calculated according to miRNA comprehensive similarity matrix  $SM$  as follows:

$$W_{ij}^m = \begin{cases} 1, & i \in N_k(m_j) \& j \in N_k(m_i) \\ 0, & i \notin N_k(m_j) \& j \notin N_k(m_i) \\ 0.5, & otherwise \end{cases} \quad (29)$$

where  $N_k(m_i)$  and  $N_k(m_j)$  denote the sets of  $k$ -nearest neighbors of miRNA  $m_i$  and miRNA  $m_j$ , respectively. Then, the graph matrix  $SM^*$  for miRNAs is presented as follows:

$$\forall i, j \quad SM_{ij}^* = SM_{ij} W_{ij}^m \quad (30)$$

Similarly, we used the same method to calculate graph matrix  $SD^*$  for disease. In this paper, we utilized final miRNA similarity network  $SM^*$ , disease similarity network  $SD^*$  and known miRNA-disease association network to construct a miRNA-disease heterogeneous network. Then, the KNN method was used to learn hypergraphs related to miRNAs and diseases based on the constructed heterogeneous network. Finally, according to learned hypergraphs, we calculated miRNA hypergraph Laplacian matrix  $L_{hm}^*$  and disease hypergraph Laplacian matrix  $L_{hd}^*$ , respectively.

Non-negative matrix tri-factorization (NMTF) has been widely applied for data representation in various fields [51,

52]. The purpose of NMTF is to obtain three low-dimensional non-negative factor matrices for low-dimensional approximation of the original matrix. However, the bad local optimal solutions are often encountered when solving non-convex optimized NMTF. To effectively alleviate the model from falling into a bad local optimal solution, we combined self-paced learning with NMTF. Moreover, we introduced hypergraph regularization to better preserve the high-order relations of heterogeneous network in low-dimensional space. In addition,  $L_{2,1}$  norm was utilized to replace Frobenius norm for calculating residual error, thus reducing the impact of noise and outliers on prediction performance. Finally, the objective function was defined as:

$$\begin{aligned} \min_{U, S, V, w} & \|diag(w)(X - USV^T)\|_{2,1} + \lambda \|S\|_F^2 \\ & + \alpha Tr(U^T L_{hm}^* U) + \beta Tr(V^T L_{hd}^* V) + g(\eta, w) \\ \text{s.t. } & U \geq 0, S \geq 0, V \geq 0, U^T U = I, V^T V = I \end{aligned} \quad (31)$$

where  $\alpha$  and  $\beta$  control the importance of hypergraph Laplacian regularization term.  $\odot$  represents the Hadamard product. Meanwhile, we imposed orthogonal constraints on the factor matrices  $U$  and  $V$  to enhance the uniqueness and stability of the decomposition results. To prevent overfitting,  $L_2$  norm was utilized to constrain factor matrix  $S$ , and  $\lambda$  is regularization coefficient.

For ease of calculation, Eq. (31) can be reformulated by optimizing the following problem:

$$\begin{aligned} \min_{U, S, V, w} & \sum_{i=1}^n d_i \| (X - USV^T)_i \|_2^2 + \lambda \|S\|_F^2 \\ & + \alpha Tr(U^T L_{hm}^* U) + \beta Tr(V^T L_{hd}^* V) + g(\eta, w) \\ \text{s.t. } & U \geq 0, S \geq 0, V \geq 0, U^T U = I, V^T V = I \end{aligned} \quad (32)$$

where

$$d_i = \frac{w_i}{2 \| (X - USV^T)_i \|_2} \quad (33)$$

With simple algebra, Eq. (32) can be written as follows:

$$\begin{aligned} \min_{U, S, V, w} & Tr((X - USV^T)D(X - USV^T)^T) + \lambda Tr(SS^T) \\ & + \alpha Tr(U^T L_{hm}^* U) + \beta Tr(V^T L_{hd}^* V) + g(\eta, w) \\ \text{s.t. } & U \geq 0, S \geq 0, V \geq 0, U^T U = I, V^T V = I \end{aligned} \quad (34)$$

where  $D$  is a diagonal matrix and  $D_{ii} = d_i$ .

In this work, the alternate iterative updating algorithm was proposed to solve the optimization problem of SPLHRNMTF. More specifically, the objective function alternately optimizes one variable while fixing other variables.



**Updating the weight variable  $w$**

When  $U$ ,  $S$  and  $V$  were fixed, optimizing the subproblem involving  $w$  was as follows:

$$\min_w \sum_{i=1}^n w_i l_i + g(\eta, w) \tag{35}$$

where  $l_i = \|(X - USV^T)_i\|_2$ , which represents the reconstruction error of the  $i$ -th sample.

On the basis of previous work [45], a novel soft weighting SPL regularization term was utilized as follows:

$$g(\eta, w) = - \sum_{i=1}^n \zeta \ln(w_i + \zeta/\eta) \tag{36}$$

Clearly, the optimal  $w^*$  can be easily calculated by

$$w_i^* = \begin{cases} 1, & \text{if } l_i \leq \zeta\eta/(\zeta + \eta) \\ 0, & \text{if } l_i \geq \eta \\ \zeta/l_i - \zeta/\eta, & \text{otherwise.} \end{cases} \tag{37}$$

**Updating the factor matrix  $U$**

According to the strategy of alternate updating, when the other variables were fixed, the terms in the objective function involving  $U$  can be reformulated as follows:

$$\begin{aligned} \min_U & Tr(XDX^T) - 2Tr(USV^TDX^T) \\ & + Tr(USV^TDVSTU^T) + \alpha Tr(U^TL_{hm}^*U) \tag{38} \\ \text{s.t. } & U \geq 0, U^TU = I \end{aligned}$$

Let  $\Phi$  and  $\Psi$  be the Lagrange multiplier for constraints  $U \geq 0$  and  $U^TU = I$ , respectively. Then, the Lagrange function is

$$\begin{aligned} \mathcal{L}_1 = & Tr(XDX^T) - 2Tr(USV^TDX^T) \\ & + Tr(USV^TDVSTU^T) + \alpha Tr(U^TL_{hm}^*U) \tag{39} \\ & + Tr(\Phi U^T) + Tr(\Psi(U^TU - I)^T) \end{aligned}$$

The partial derivative with respect to the factor matrix  $U$  can be obtained as follows:

$$\begin{aligned} \frac{\partial \mathcal{L}_1}{\partial U} = & -2XDVST + 2USV^TDVST \\ & + 2\alpha L_{hm}^*U + 2U\Psi + \Phi \tag{40} \end{aligned}$$

Setting  $\frac{\partial \mathcal{L}_1}{\partial U} = 0$ , substituting  $\Psi = U^TXDVST - SV^TDVST - \alpha U^TL_{hm}^*U$  and using the Karush-Kuhn-Tucker (KKT) conditions  $\Phi_{ij}U_{ij} = 0$ , we can obtain the following update rule:

$$U_{ij} \leftarrow U_{ij} \sqrt{\frac{(XDVST + \alpha S_{hm}^*U)_{ij}}{(UU^TXDVST + \alpha UU^TL_{hm}^*U)_{ij}}} \tag{41}$$

**Updating the factor matrix  $S$**

When  $w$ ,  $U$  and  $V$  were fixed, the subproblem of optimizing an objective function containing  $S$  was as below:

$$\begin{aligned} \min_S & Tr(XDX^T) - 2Tr(USV^TDX^T) \\ & + Tr(USV^TDVSTU^T) + \lambda Tr(SS^T) \tag{42} \\ \text{s.t. } & S \geq 0 \end{aligned}$$

Let  $\Xi$  be the Lagrange multiplier for constraint  $S \geq 0$ . Then, the Lagrange function is

$$\begin{aligned} \mathcal{L}_2 = & Tr(XDX^T) - 2Tr(USV^TDX^T) + \lambda Tr(SS^T) \\ & + Tr(USV^TDVSTU^T) + Tr(\Xi S^T) \tag{43} \end{aligned}$$

The partial derivative of the factor matrix  $S$  can be calculated as below:

$$\frac{\partial \mathcal{L}_2}{\partial S} = -2U^TXDV + 2U^TUSV^TDV + 2\lambda S + \Xi \tag{44}$$

Setting  $\frac{\partial \mathcal{L}_2}{\partial S} = 0$  and using KKT conditions  $\Xi_{ij}S_{ij} = 0$ , we obtained the following update rule:

$$S_{ij} \leftarrow S_{ij} \sqrt{\frac{(U^TXDV)_{ij}}{(U^TUSV^TDV + \lambda S)_{ij}}} \tag{45}$$

**Updating the factor matrix  $V$**

When  $w$ ,  $U$  and  $S$  were fixed, optimizing the subproblem involving  $V$  was as follows:

$$\begin{aligned} \min_V & Tr(XDX^T) - 2Tr(USV^TDX^T) \\ & + Tr(USV^TDVSTU^T) + \beta Tr(V^TL_{hd}^*V) \tag{46} \\ \text{s.t. } & V \geq 0, V^TV = I \end{aligned}$$

Let  $\Theta$  and  $\Lambda$  be the Lagrange multiplier for constraints  $V \geq 0$  and  $V^TV = I$ , respectively. Then, the Lagrange function is

$$\begin{aligned} \mathcal{L}_3 = & Tr(XDX^T) - 2Tr(USV^TDX^T) \\ & + Tr(USV^TDVSTU^T) + \beta Tr(V^TL_{hd}^*V) \tag{47} \\ & + Tr(\Theta V^T) + Tr(\Lambda(V^TV - I)^T) \end{aligned}$$

The partial derivative with respect to the factor matrix  $V$  can be calculated as follows:

$$\frac{\partial \mathcal{L}_3}{\partial V} = -2DX^TUS + 2DVS^T U^T US + 2\beta L_{hd}^* V + 2V\Lambda + \Theta \quad (48)$$

Similarly, setting  $\frac{\partial \mathcal{L}_3}{\partial V} = 0$ , substituting  $\Lambda = V^TDX^TUS - V^TDVS^T U^T US - \beta V^TL_{hd}^* V$  and using KKT

conditions  $\Theta_{ij}V_{ij} = 0$ , we obtained the following update rule:

$$V_{ij} \leftarrow V_{ij} \sqrt{\frac{(DX^TUS + \beta S_{hd}^* V)_{ij}}{(DVV^T X^T US + \beta VV^T S_{hd}^* V)_{ij}}} \quad (49)$$

The detailed optimization of the proposed SPLHRN-MTF model was described step by step in Algorithm 1.

**Algorithm 1** The optimization algorithm of SPLHRNMTF

---

**Input:** known miRNA-disease association matrix  $X$ ; comprehensive similarity for miRNAs  $SM$ ; comprehensive similarity for diseases  $SD$ ;  $\alpha$ ,  $\beta$  and  $\lambda$ .

**Output:** Predicted association matrix  $\hat{X}$ .

- 1: Randomly initialize non-negative matrices  $U$ ,  $S$  and  $V$ .
- 2: **for** each miRNA  $m_p \in M = \{m_1, m_2, \dots, m_m\}$  **do**
- 3:      $X_m(m_p) = \frac{1}{E_m} \sum_{i=1}^K \theta^{i-1} \times SM(m_i, m_p) \times X(m_i)$
- 4: **end for**
- 5: **for** each disease  $d_q \in D = \{d_1, d_2, \dots, d_n\}$  **do**
- 6:      $X_d(d_q) = \frac{1}{E_d} \sum_{j=1}^K \theta^{j-1} \times SD(d_j, d_q) \times X(d_j)$
- 7: **end for**
- 8: Merge the new interaction profiles  $X_{md} = (X_m + X_d)/2$ .
- 9: Update association matrix  $X = \max(X, X_{md})$ .
- 10: Construct sparse similarity matrices  $SM^*$ ,  $SD^*$  based on  $SM$ ,  $SD$ .
- 11: **repeat**
- 12:     Fix  $U$ ,  $S$  and  $V$ , update  $w$  using Eq. (37)
- 13:     **repeat**
- 14:         Fix  $w$ , update  $U$ ,  $S$  and  $V$
- 15:         Update  $U$  using Eq. (41)
- 16:         Update  $S$  using Eq. (45)
- 17:         Update  $V$  using Eq. (49)
- 18:     **until** convergence
- 19: **until** All the instances are chosen
- 20: **return** Model parameters  $U$ ,  $S$  and  $V$ .

---

### Convergence analysis

In this section, we will study the convergence of the SPLHRNMTF model. To be more specific, we needed to prove that the objective function in Eq. (31) is monotonically decreasing under the updating rules in Eqs. (41), (45) and (49) and the SPL step in Eq. (37).

Based on previous works [43, 53, 54], researchers have demonstrated the effectiveness of SPL in various models. Finally, we only needed to prove that the objective function in Eq. (31) is monotonically decreasing under the updating rules in Eqs. (41), (45) and (49). The following detailed proof process was presented.

**Definition 1**  $G(x, x')$  is an auxiliary function for  $F(x)$  if the conditions

$$G(x, x') \geq F(x), G(x, x) = F(x)$$

are satisfied.

*lemma 1* If  $G(x, x')$  is an auxiliary function of  $F(x)$ , then  $F(x)$  is non-increasing under the following updating formula:

$$x^{t+1} = \arg \min_x G(x, x^t)$$

**Proof**

$$F(x^{t+1}) \leq G(x^{t+1}, x^t) \leq G(x^t, x^t) = F(x^t)$$

Then,

$$F(x^{min}) \leq \dots \leq F(x^{t+1}) \leq F(x^t) \leq \dots \leq F(x^0)$$

The above mentioned is to introduce the definition of the auxiliary function. Next, in order to demonstrate the convergence of the alternating optimization rules, we needed the following theorems.

**Theorem 1** Updating  $U$  using Eq. (41) while fixing  $S$  and  $V$  in each iteration will monotonically decrease the value of the objective function in Eq. (31), hence it converges.

According to Eq. (31), the corresponding objective function when fixed  $w$  can be rewritten as below:

$$\begin{aligned} \mathcal{J}(U, S, V) = & \sum_{i=1}^n d_i \|X - USV^T\|_2^2 + \lambda \|S\|_F^2 \\ & + \alpha \text{Tr}(U^T L_{hm}^* U) + \beta \text{Tr}(V^T L_{hd}^* V) \\ \text{s.t. } & U \geq 0, S \geq 0, V \geq 0, U^T U = I, V^T V = I \end{aligned} \quad (50)$$

Our goal is to prove that the updating rule for  $U$  while fixing  $S$  and  $V$  will result in a monotonically decreasing value of  $\mathcal{J}(U, S, V)$ . To conduct the corresponding proof, we needed the following lemma.

*lemma 2* Updating  $U$  using Eq. (41) while fixing  $S$  and  $V$ , the following inequality holds:

$$\begin{aligned} & \sum_{i=1}^n d_i \|X - U^{t+1} S V^T\|_2^2 + \alpha \text{Tr}(U^{T(t+1)} L_{hm}^* U^{(t+1)}) \leq \\ & \sum_{i=1}^n d_i \|X - U^t S V^T\|_2^2 + \alpha \text{Tr}(U^{T(t)} L_{hm}^* U^{(t)}) \end{aligned} \quad (51)$$

where  $U^T U = I$ .

**Proof** To prove the constrained problem in Lemma 2, we needed to prove the following Lagrangian function:

$$\begin{aligned} \mathcal{J}(U) = & \text{Tr}(XDX^T) - 2\text{Tr}(USV^TDX^T) \\ & + \text{Tr}(USV^TDVS^TU^T) + \alpha \text{Tr}(U^T L_{hm}^* U) \\ & + \text{Tr}(\Psi(U^T U - I)^T) \end{aligned} \quad (52)$$

The auxiliary function can be first defined as follows:

$$\begin{aligned} \mathcal{J}(U, U') = & \text{Tr}(XDX^T - \Psi) \\ & + \sum_{j=1}^c \sum_{i=1}^n \frac{(U' S V^T D V S^T + U' \Psi)_{ij} U'_{ij}^2}{U'_{ij}} \\ & - 2 \sum_{j=1}^c \sum_{i=1}^n (X D V S^T)_{ij} U'_{ij} \left( 1 + \log \frac{U'_{ij}}{U'_{ij}} \right) \\ & + \alpha \sum_{j=1}^c \sum_{i=1}^n \frac{(L_{hm}^* U')_{ij} U'_{ij}^2}{U'_{ij}} \end{aligned} \quad (53)$$

According to previous work [55], the following inequality holds:

$$\sum_{i=1}^n \sum_{j=1}^k \frac{(AC'B)_{ij} C_{ij}^2}{C_{ij}} \geq \text{Tr}(C^T ACB) \quad (54)$$

where  $A, B$  and  $C$  are non-negative matrices, and  $A$  and  $B$  are symmetric. Furthermore, we knew the inequality,  $w \geq 1 + \log(w)$  for all  $w > 0$ . Therefore, we had

$$\begin{aligned} & \text{Tr}(U^T USV^TDVS^TU) + \text{Tr}(\Psi U^T U) \\ & \leq \sum_{j=1}^c \sum_{i=1}^n \frac{(U' S V^T D V S^T)_{ij} U'_{ij}^2}{U'_{ij}} + \sum_{j=1}^c \sum_{i=1}^n \frac{(U' \Psi)_{ij} U'_{ij}^2}{U'_{ij}} \\ & = \sum_{j=1}^c \sum_{i=1}^n \frac{(U' S V^T D V S^T + U' \Psi)_{ij} U'_{ij}^2}{U'_{ij}} \end{aligned} \quad (55)$$

and

$$\sum_{j=1}^c \sum_{i=1}^n \frac{(L_{hm}^* U')_{ij} U'_{ij}^2}{U'_{ij}} \geq \text{Tr}(U'^T L_{hm}^* U') \quad (56)$$

It can be seen that Eq. (53) can serve as a valid auxiliary function. Consequently, we can determine the stationary point of  $\mathcal{J}(U, U')$ . As defined in Eq. (53), taking the derivative of  $\mathcal{J}(U, U')$  w.r.t  $U$

$$\frac{\partial \mathcal{J}(U, U')}{\partial U_{ij}} = 2 \frac{(U' S V^T D V S^T + \alpha L_{hm}^* U' + U' \Psi)_{ij} U_{ij}}{U'_{ij}} - 2(X D V S^T)_{ij} \frac{U'_{ij}}{U_{ij}} \tag{57}$$

Setting Eq. (57) to zero and substituting  $\Psi = U'^T X D V S^T - S V^T D V S^T - \alpha U'^T L_{hm}^* U'$ , we can obtain the following stationary point:

$$U_{ij} \leftarrow U'_{ij} \sqrt{\frac{(X D V S^T + \alpha S_{hm}^* U')_{ij}}{(U' U'^T X D V S^T + \alpha U' U'^T S_{hm}^* U')_{ij}}} \tag{58}$$

To confirm that the stationary point is the minimum of  $\mathcal{J}(U, U')$ , it is necessary to examine whether the Hessian matrix is a positive semidefinite matrix. To this end, we further took the second derivative w.r.t  $U$

$$\frac{\partial^2 \mathcal{J}(U, U')}{\partial V_{ij} \partial V_{kl}} = \left( 2(X D V S^T)_{ij} \frac{U'_{ij}}{U_{ij}} \right) \delta_{ik} \delta_{jl} + \left( 2 \frac{(U' S V^T D V S^T + \alpha L_{hm}^* U' + U' \Psi)_{ij} U_{ij}}{U'_{ij}} \right) \delta_{ik} \delta_{jl} \tag{59}$$

It is clear that the Hessian matrix is a positive semidefinite matrix, thereby indicating  $\mathcal{J}(U, U')$  is a convex function. This observation suggests that the stationary point in Eq. (58) represents the unique global minima of  $\mathcal{J}(U, U')$ . Based on Lemma 1, it becomes evident that Lemma 2 is established. Hence, Theorem 1 has been proven too. It is worth noting that if substituting  $U = U^{t+1}$  and  $U' = U^t$  for Eq. (58), we obtained the updating rule in Eq. (41).

**Theorem 2** Updating  $S$  using Eq. (45) while fixing  $U$  and  $V$  in each iteration will monotonically decrease the value of the objective function in Eq. (31), hence it converges.

In the same way, our goal is to prove that the updating rule for  $S$  while keeping  $U$  and  $V$  fixed, results in a monotonically decreasing value of  $\mathcal{J}(U, S, V)$ . To conduct the corresponding proof, we needed the lemma as follows.

*lemma 3* Updating  $S$  using Eq. (45) while fixing  $U$  and  $V$ , the following inequality holds:

$$\sum_{i=1}^n d_i \| (X - U S^{t+1} V^T)_i \|_2^2 + \lambda \| S^{t+1} \|_F^2 \leq \sum_{i=1}^n d_i \| (X - U S^t V^T)_i \|_2^2 + \lambda \| S^t \|_F^2 \tag{60}$$

**Proof** Similar with Eq. (52), we can get the Lagrangian function for  $\mathcal{J}(S)$  as follows:

$$\mathcal{J}(S) = Tr(X D X^T) - 2Tr(USV^T D X^T) + Tr(USV^T D V S^T U^T) + \lambda Tr(SS^T) \tag{61}$$

Also, an appropriate auxiliary function was defined as:

$$\begin{aligned} \mathcal{J}(S, S') &= Tr(X D X^T) \\ &+ \sum_{j=1}^c \sum_{i=1}^n \frac{(U^T U S' V^T D V)_{ij} S'_{ij}}{S'_{ij}} \\ &- 2 \sum_{j=1}^c \sum_{i=1}^n (U^T X D V)_{ij} S'_{ij} \left( 1 + \log \frac{S'_{ij}}{S_{ij}} \right) \\ &+ \lambda \sum_{j=1}^c \sum_{i=1}^n \frac{(S')_{ij} S'_{ij}}{S'_{ij}} \end{aligned} \tag{62}$$

Similar to prove Lemma 2, Lemma 3 can be proved. The rest proof of Theorem 2 is similar to the proof of Theorem 1.

**Theorem 3** Updating  $V$  using Eq. (49) while fixing  $U$  and  $S$  in each iteration will monotonically decrease the value of the objective function in Eq. (31), hence it converges.

Similarly, our goal is to prove that the updating rule for  $V$  while keeping  $U$  and  $S$  fixed, results in a monotonically decreasing value of  $\mathcal{J}(U, S, V)$ . To conduct the corresponding proof, we needed the lemma as follows.

*lemma 4* Updating  $V$  using Eq. (49) while fixing  $U$  and  $S$ , the following inequality holds:

$$\begin{aligned} \sum_{i=1}^n d_i \| (X - U S V^{t+1})_i \|_2^2 + \beta Tr(V^{T(t+1)} L_{hd}^* V^{t+1}) &\leq \\ \sum_{i=1}^n d_i \| (X - U S V^t)_i \|_2^2 + \beta Tr(V^{T(t)} L_{hd}^* V^t) \end{aligned} \tag{63}$$

where  $V^T V = I$ .

**Proof** Similar with Eq. (52), we obtained the Lagrangian function for  $\mathcal{J}(V)$  and an appropriate auxiliary function as follows:

$$\begin{aligned} \mathcal{J}(V) &= Tr(X D X^T) - 2Tr(USV^T D X^T) \\ &+ Tr(USV^T D V S^T U^T) + \beta Tr(V^T L_{hd}^* V) \\ &+ Tr(\Lambda(V^T V - I)^T) \end{aligned} \tag{64}$$

$$\begin{aligned}
 \mathcal{J}(V, V') &= Tr(XDX^T - \Lambda) \\
 &+ \sum_{j=1}^c \sum_{i=1}^n \frac{(DV'S^T U^T US + V'\Lambda)_{ij} U_{ij}^2}{U_{ij}'} \\
 &- 2 \sum_{j=1}^c \sum_{i=1}^n (DX^T US)_{ij} V'_{ij} \left(1 + \log \frac{V_{ij}}{V'_{ij}}\right) \\
 &+ \beta \sum_{j=1}^c \sum_{i=1}^n \frac{(L_{hd}^* V')_{ij} V_{ij}^2}{V'_{ij}}
 \end{aligned}
 \tag{65}$$

Similar to prove Lemma 2, we can prove Lemma 4. The rest proof of Theorem 3 is also similar to the proof of Theorem 1.

In summary, the objective function in Eq. (31) decreases monotonically in the alternative updating rules in Eqs. (41), (45) and (49).

## Results

### Experimental settings and evaluation metrics

To more comprehensively and systematically show the ability of SPLHRNMTF in predicting potential associations between miRNAs and diseases, we conducted 5-fold cross-validation five times experiments based on known miRNA-disease associations. In 5-fold cross-validation, we randomly split all miRNA-disease pairs into five equal-sized subsets. Note that there is no overlap between five equal-sized subsets. In each fold, one subset as a testing set in turn and the remaining four subsets are considered as a training set. To effectively evaluate the prediction ability of the model, the area under the receiver operating characteristic (AUC) curve, the area under the precision-recall (AUPR) curve, and F1 score were used to evaluate the performance of all models.

### Baseline models

To comprehensively evaluate the prediction performance of SPLHRNMTF, we compared several previously proposed computational models. In this paper, we applied the same dataset to train these models for a fairer comparative analysis. The specific information of introduced models was as follows.

NMTF [56]: As an extension of NMF, non-negative matrix tri-factorization (NMTF) can provide more degrees of freedom than NMF.

$L_{2,1}$ -NMTF [57]: Kong et al. proposed robust non-negative matrix factorization by using  $L_{2,1}$  norm instead of Frobenius norm. Based on the study of Kong et al., we extended NMF to NMTF and named it  $L_{2,1}$ -NMTF.

GRNMF [22]: This is a method based on graph regularized non-negative matrix factorization for miRNA-disease association prediction.

IMCMDA [23]: Chen et al. developed an inductive matrix completion method based on non-negative constraints to predict the associations between miRNAs and diseases, which combines integrated miRNA and disease similarity matrices into matrix factorization.

MDHGI [24]: The matrix decomposition-based MDHGI combines sparse learning method with heterogeneous graph inference method to infer the associations between miRNAs and diseases.

SPLGRNMF [45]: Huang et al. proposed the SPLNMF model by integrating self-paced learning (SPL) into NMF and using  $L_{2,1}$  norm to constrain the objective function. To conduct a more reasonable comparative analysis with the proposed model, we imposed double orthogonal constraints on the latent representations and introduced dual graph Laplacian regularization based on the SPL-NMF model. The improved model is called SPLGRNMF.

NMTFDR [25]: Dissez et al. presented a drug repositioning model based on the non-negative matrix tri-factorization with graph Laplacian regularization by using multiple interaction matrices of drugs, proteins, pathways and diseases. For convenience, the drug repositioning model is called NMTFDR.

NMTF-DTI [26]: NMTF-DTI is a model based on non-negative matrix tri-factorization for drug-target interaction prediction, which constructs a weighted interaction matrix to correct false-negative samples and utilizes graph Laplacian regularization to improve prediction performance.

SPLNMTF [58]: SPLNMTF combines self-paced learning with non-negative matrix tri-factorization to perform drug repositioning, which also considers graph Laplacian regularization to preserve local geometric structures in low-dimensional space.

NMFMC [27]: The method divides non-negative matrix factorization into a known part and an unknown part, and introduces graph regularization to preserve local geometric structures for miRNA-disease association prediction.

NIMCGCN [19]: NIMCGCN makes use of a neural inductive matrix completion method based on graph convolutional networks to identify the associations between miRNAs and diseases.

MMGCN [20]: Tang et al. applied graph convolutional networks and multichannel attention mechanism to extract and enhance latent representations of miRNAs and diseases, predicting potential associations from the reconstructed miRNA-disease matrix.

MHCLMDA [21]: MHCLMDA first captured high-order interactions between similarity network nodes

based on hypergraph convolutional networks and hypergraph contrastive learning. Then, a variational auto-encoder was used to extract miRNA and disease features from known miRNA-disease associations. Finally, MHCLMDA integrated different miRNA and disease features to predict miRNA-disease associations.

**Parameters analysis**

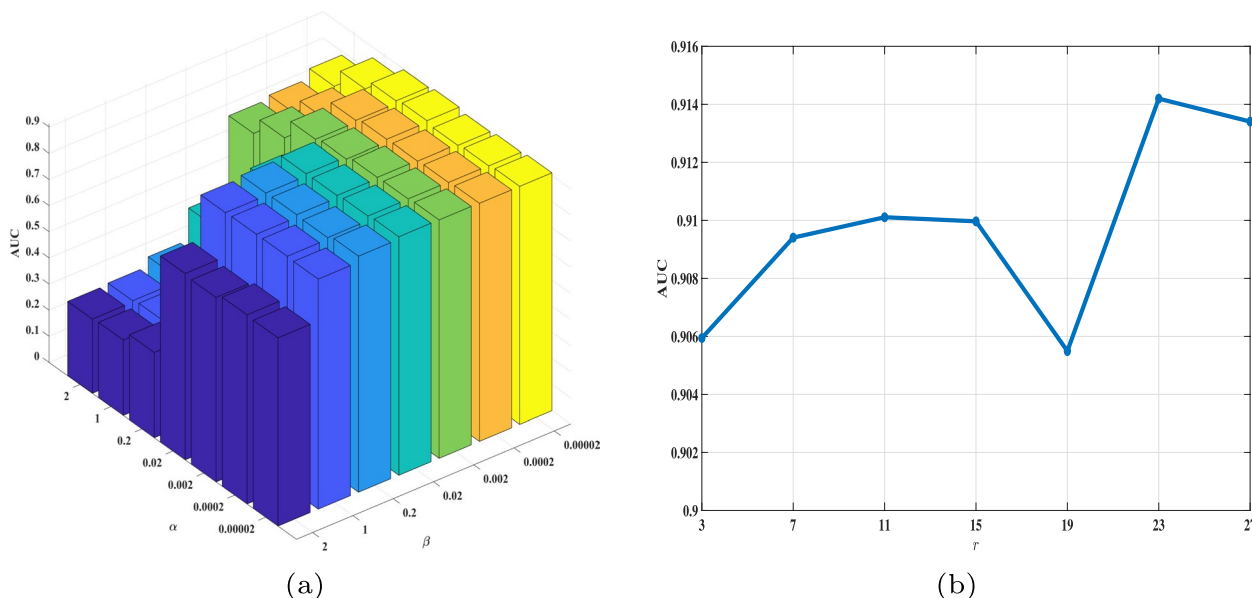
In this study, we analyzed the impact of several hyperparameters on the prediction performance of SPLHRNMTF by utilizing the MDAv2.0\_data dataset. To be more specific, we applied grid search to determine hyperparameter combinations under 5-fold cross-validation. Among them, compared with other hyperparameters,  $\lambda$  has less influence on model prediction performance. First, we empirically fixed  $\lambda = 0.0001$  and searched the optimal values of other hyperparameters. Then, we searched the optimal value of  $\alpha$  and  $\beta$  from  $\{0.00002, 0.0002, 0.002, 0.02, 0.2, 1, 2\}$ . As shown in Fig. 3a, when  $\alpha = 0.00002$  and  $\beta = 0.02$ , SPLHRNMTF obtained better prediction performance. Similarly, varying matrix rank  $r$  within  $\{3, 7, \dots, 23, 27\}$  and set  $r = 23$  in Fig. 3b. Next, the optimal  $k_1$  value in KNN method used by constructing hypergraphs can be searched from  $\{2, 6, \dots, 38, 42\}$  and set  $k_1 = 34$  as shown in Supplementary Figure S1(a). In addition, we searched the optimal decay term  $\theta$  from  $\{0.1, 0.2, 0.3, \dots, 0.9, 1\}$  and set  $\theta = 0.9$  in Supplementary Figure S1(b). The  $k_2$  in WKNN method was also searched from  $\{2, 4, 6, \dots, 12, 14\}$  and  $k_2$  was set to 10 that the model achieved better prediction

performance as shown in Supplementary Figure S1(c). Finally, we trained  $k_3$  in KNN method used by sparse similarity matrix from  $\{1, 2, 3, \dots, 9, 10\}$  and set the optimal  $k_3$  to 2 as shown in Supplementary Figure S1(d). It is worth noting that hyperparameter selection was required for other experimental datasets as well. The detailed results of hyperparameter adjustments can be found in Supplementary Figures S2 and S3.

**Comparison experiments**

All comparison experiments adopt the same datasets and evaluation settings for a fairer comparative analysis. To better show the optimal performance of baseline models, we performed grid search on GRNMF, SPLGRNMF, NMTF-DTI and NMFMC models with hyperparameters controlling the graph Laplacian regularization. The detailed hyperparameter tuning can be found in Supplementary Parameters Analysis. For fairness, we uniformly set the same rank as SPLHRNMTF for all baseline models on different datasets. To conduct more reasonable evaluation experiments, we performed 5-fold cross-validation five times. In addition, we also randomly selected unobserved elements equal to the positive sample size as negative samples 10 times. The reported average results represent the outcome, providing a more reasonable evaluation. Finally, we provided detailed parameter information of each model on three datasets in Supplementary Material.

Table 1 shows the comparative analysis of the prediction performance of all models for 5-fold cross-validation



**Fig. 3** The influence of different hyperparameters on SPLHRNMTF based on the MDAv2.0\_data dataset. **a** The impact of hyperparameters  $\alpha$  and  $\beta$  on SPLHRNMTF. **b** The impact of hyperparameter  $r$  on SPLHRNMTF

**Table 1** The prediction performance of all models evaluated by 5-fold cross-validation five times on MDAv3.2\_data1 and MDAv3.2\_data2 datasets

		AUC	AUPR	F1
MDAv3.2_data1	NMTF	0.9110±0.0006	0.9177±0.0005	0.8553±0.0004
	$L_{2,1}$ -NMTF	0.9128±0.0006	0.9193±0.0005	0.8580±0.0002
	IMCMDA	0.7783±0.0003	0.7912±0.0002	0.7238±0.0006
	GRNMF	0.9041±0.0009	0.9067±0.0010	0.8303±0.0011
	MDHGI	0.8877±0.0008	0.9096±0.0006	0.8446±0.0011
	SPLGRNMF	0.7691±0.0045	0.7294±0.0042	0.7732±0.0040
	NMTFDR	0.9073±0.0014	0.9157±0.0011	0.8419±0.0021
	NMTF-DTI	0.9279±0.0004	0.9340±0.0002	0.8602±0.0005
	SPLNMTF	0.8428±0.0003	0.8325±0.0001	0.7935±0.0008
	NMFMC	0.9082±0.0003	0.9018±0.0002	0.8382±0.0009
	NIMCGCN	0.8540±0.0009	0.8578±0.0010	0.7825±0.0004
	MMGCN	0.9004±0.0018	0.9213±0.0008	0.8575±0.0011
	MHCLMDA	0.9255±0.0009	0.9347±0.0012	0.8591±0.0008
	SPLHRNMTF	<b>0.9304±0.0002</b>	<b>0.9349±0.0002</b>	<b>0.8614±0.0004</b>
MDAv3.2_data2	NMTF	0.8096±0.0020	0.8153±0.0019	0.7600±0.0018
	$L_{2,1}$ -NMTF	0.8425±0.0042	0.8504±0.0041	0.7709±0.0012
	IMCMDA	0.8406±0.0003	0.8512±0.0001	0.7849±0.0005
	GRNMF	0.9170±0.0004	0.9248±0.0005	0.8442±0.0012
	MDHGI	0.9011±0.0008	0.9195±0.0006	0.8545±0.0017
	SPLGRNMF	0.7682±0.0048	0.7524±0.0032	0.7544±0.0035
	NMTFDR	0.9149±0.0014	0.9204±0.0011	0.8491±0.0017
	NMTF-DTI	0.9316±0.0006	0.9356±0.0004	0.8620±0.0006
	SPLNMTF	0.9311±0.0004	0.9353±0.0003	0.8629±0.0004
	NMFMC	0.9069±0.0005	0.8957±0.0004	0.8405±0.0012
	NIMCGCN	0.9057±0.0004	0.9075±0.0006	0.8378±0.0003
	MMGCN	0.9179±0.0016	0.9367±0.0009	<b>0.8706±0.0012</b>
	MHCLMDA	0.9313±0.0012	0.9379±0.0019	0.8678±0.0010
	SPLHRNMTF	<b>0.9358±0.0004</b>	<b>0.9387±0.0003</b>	0.8642±0.0006

five times, from which we can see that SPLHRNMTF outperformed other baseline models, even including graph and hypergraph convolutional network-based models, on MDAv3.2\_data1 and MDAv3.2\_data2 datasets in terms of AUC and AUPR. To be more specific, the average AUC value of SPLHRNMTF was 0.9304 on the MDAv3.2\_data1 dataset, whereas the average AUC values of NMTF,  $L_{2,1}$ -NMTF, IMCMDA, GRNMF, MDHGI, SPLGRNMF, NMTFDR, NMTF-DTI, SPLNMTF, NMFMC, NIMCGCN, MMGCN and MHCLMDA were 0.9110, 0.9128, 0.7783, 0.9041, 0.8877, 0.7691, 0.9073, 0.9279, 0.8428, 0.9082, 0.8540, 0.9004, 0.9255, respectively. Similarly, SPLHRNMTF obtained the average AUC value of 0.9358, which was also better than other baseline models on the MDAv3.2\_data2 dataset. It is worth noting that the standard deviation of the prediction performance of SPLHRNMTF was relatively small. These results indicate that SPLHRNMTF not only has superior prediction performance but also exhibits better robustness

and stability. Although the F1 value of SPLHRNMTF is slightly lower than that of MMGCN and MHCLMDA on the MDAv3.2\_data2 dataset, the AUC and AUPR values of SPLHRNMTF are better than those of MMGCN and MHCLMDA. This shows that even compared to models based on graph and hypergraph convolutional networks, SPLHRNMTF has a certain competitiveness. From Supplementary Table S1, we observed that the proposed SPLHRNMTF model also obtained good prediction performance on the MDAv2.0\_data dataset. More importantly, the prediction performance of  $L_{2,1}$ -NMTF was superior to NMTF. Specifically,  $L_{2,1}$ -NMTF obtained certain performance gains over NMTF by 5.98%, 0.20% and 4.06% in terms of AUC on MDAv2.0\_data, MDAv3.2\_data1 and MDAv3.2\_data2 datasets, respectively. This result demonstrates that using  $L_{2,1}$  to calculate residual error can effectively alleviate the impact of noise and outliers, thereby improving prediction performance.

### Ablation studies

To better validate the effectiveness of non-linear fusion method, preprocessing step using WKNN, self-paced learning and hypergraph regularization, we further constructed SPLHRNMTF<sup>-</sup>, SPLHRNMTF\*, HRNMTF and SPLGRNMTF as four variants of SPLHRNMTF for comparative analysis. (1) SPLHRNMTF<sup>-</sup>: we utilized linear fusion instead of non-linear fusion to explore whether learning non-linear relationships in similarity networks can improve prediction performance. (2) SPLHRNMTF\*: we removed the preprocessing step using WKNN to explore whether using the WKNN method can effectively correct false-negative samples and improve model prediction performance. (3) HRNMTF: in order to verify the effectiveness of self-paced learning, we retained other modules except self-paced learning. (4) SPLGRNMTF: we replaced hypergraph Laplacian regularization with graph Laplacian regularization to explore whether the use of hypergraph regularization can effectively capture high-order relations.

Table 2 shows the comparative analysis of ablation experiments for different variant models on the MDAv3.2\_data2 dataset. From Table 2, we can observe that SPLHRNMTF was significantly better than SPLHRNMTF\*. This result demonstrates that the use of WKNN method can effectively correct false-negative samples, thereby improving model prediction performance. Moreover, after using non-linear fusion method, the prediction performance of SPLHRNMTF outperformed SPLHRNMTF<sup>-</sup>, which suggests that the non-linear fusion method can more reasonably integrate similarity information while improving prediction performance by capturing non-linear relationships. Similarly, after using self-paced learning, SPLHRNMTF can achieve better prediction performance than HRNMTF, which shows that integrating self-paced learning into NMTF can effectively alleviate the model from falling into a bad local optimal solution and improve prediction performance. Finally, compared with SPLGRNMTF, the prediction performance of SPLHRNMTF has been

improved to a certain extent, which indicates that the use of hypergraph regularization can capture the high-order relationships of biological similarity networks. Notably, the standard deviation of SPLHRNMTF's prediction performance in the 5-fold cross-validation five times is relatively small, indicating that adding other modules helps to enhance the model's robustness and stability.

### Case studies

To further validate the predictive accuracy of SPLHRNMTF in identifying associations between miRNAs and specific diseases, we conducted case studies on two significant tumor diseases, namely breast neoplasms and lung neoplasms, using the MDAv2.0\_data dataset. For a specific disease, negative miRNA-disease associations and experimentally verified positive miRNA-disease associations on the MDAv2.0\_data dataset were used as training samples, and unverified associations with the specific disease on the MDAv2.0\_data dataset are considered as candidate samples. Through training the SPLHRNMTF model on training samples, we can rank candidate samples by their predicted association scores and select the top 50 candidate associations with the specific disease. Finally, we verified the top 50 prediction results by finding supporting evidence from the lasted HMDD v4.0 [59] and dbDEMC [60].

Table 3 shows the top 50 prediction results of miRNAs that are closely related to breast neoplasms. From Table 3, we can see that 48 of the top 50 breast neoplasms-related miRNAs were successfully confirmed by HMDD v4.0 and dbDEMC databases. Note that unconfirmed refers to miRNAs that have not been confirmed by relevant evidence. Similarly, Supplementary Table S2 demonstrates that the top 50 predicted lung neoplasms-related results and 46 predictions can be confirmed according to the above two databases. Meanwhile, we also observed that miRNAs showing higher similarity were predicted to be associated with the same specific disease. For example, hsa-mir-138-1 and hsa-mir-138-2 can function by directly targeting the polycomb epigenetic regulator EZH2, and they have been demonstrated to be novel regulators of invasion and epithelial-mesenchymal transition in breast cancer cells [61].

To further validate the biological significance of the potential miRNA-disease associations identified by the SPLHRNMTF model, we conducted enrichment analysis on gene sets comprising the target genes of the miRNAs. First, miRTarBase [62] was used to obtain the target genes of each miRNA. Then, we applied Metascape [63] to investigate the biological process and pathway information related to these target gene sets. Figure 4 shows the enrichment analysis of the discovered breast neoplasms-related hsa-mir-19b-2, from

**Table 2** The prediction performance of ablation experiment evaluated by 5-fold cross-validation five times on the MDAv3.2\_data2 dataset

	AUC	AUPR	F1
HRNMTF	0.9298±0.0005	0.9332±0.0003	0.8609±0.0008
SPLGRNMTF	0.9318±0.0006	0.9359±0.0005	0.8623±0.0011
SPLHRNMTF <sup>-</sup>	0.9081±0.0039	0.9221±0.0028	0.8470±0.0054
SPLHRNMTF*	0.8890±0.0106	0.9069±0.0088	0.8296±0.0106
SPLHRNMTF	<b>0.9358±0.0004</b>	<b>0.9387±0.0003</b>	<b>0.8642±0.0006</b>



**Table 3** Top 50 breast neoplasms-related miRNAs predicted by SPLHRNMTF based on the MDAv2.0\_data dataset. Note that the number in evidence means PubMed Unique Identifier (PMID)

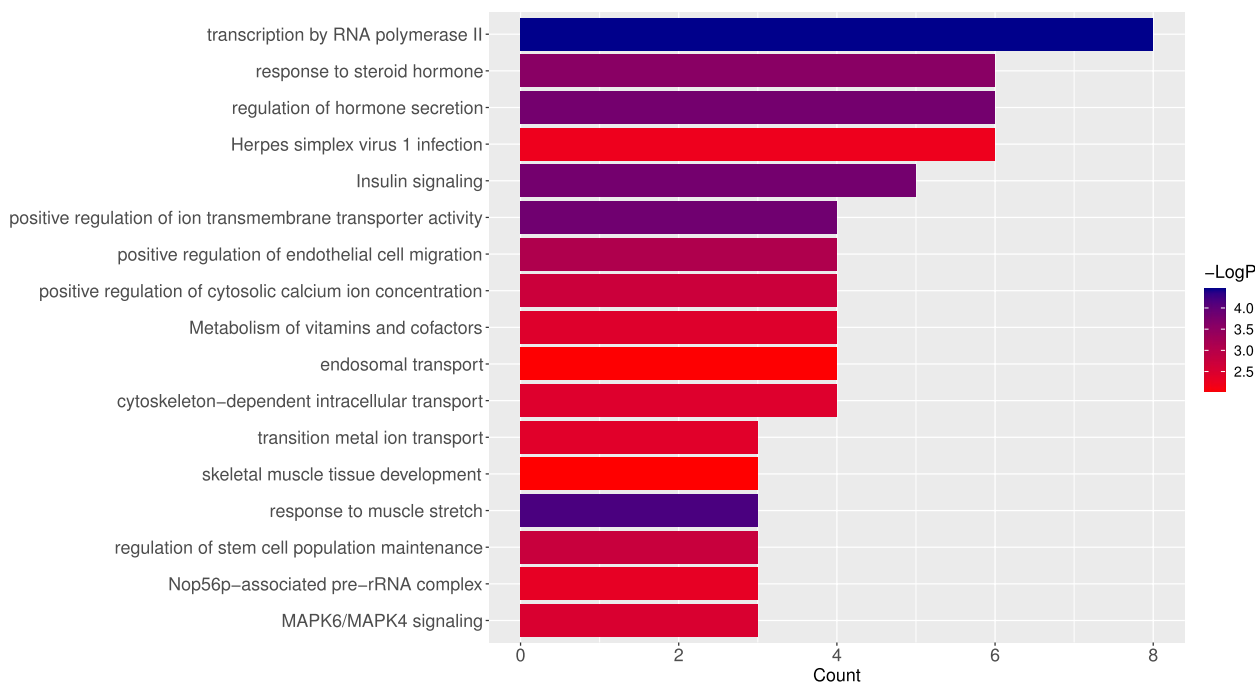
Rank	miRNA	Evidence	Rank	miRNA	Evidence
1	hsa-mir-142	33962174	26	hsa-mir-95	dbDEMCM
2	hsa-mir-106a	34837907	27	hsa-mir-32	33638154
3	hsa-mir-130a	35797350	28	hsa-mir-144	35551606
4	hsa-mir-99a	33177704	29	hsa-mir-92b	32143670
5	hsa-mir-138-2	dbDEMCM	30	hsa-mir-449b	32374522
6	hsa-mir-150	30628646	31	hsa-mir-181d	31955007
7	hsa-mir-19b-2	unconfirmed	32	hsa-mir-30e	dbDEMCM
8	hsa-mir-138-1	dbDEMCM	33	hsa-mir-134	34378285
9	hsa-mir-192	31485620	34	hsa-mir-483	33446575
10	hsa-mir-15b	32016977	35	hsa-mir-370	33629796
11	hsa-mir-196b	30416655	36	hsa-mir-424	35666424
12	hsa-mir-186	35351581	37	hsa-mir-494	32372307
13	hsa-mir-378a	34113564	38	hsa-mir-28	34593318
14	hsa-mir-449a	32952657	39	hsa-mir-184	34000513
15	hsa-mir-185	32366289	40	hsa-mir-491	33488122
16	hsa-mir-130b	34434057	41	hsa-mir-376a-1	unconfirmed
17	hsa-mir-98	34837929	42	hsa-mir-376a-2	dbDEMCM
18	hsa-mir-574	33989902	43	hsa-mir-216a	32230799
19	hsa-mir-99b	36381329	44	hsa-mir-361	34101749
20	hsa-mir-542	35318311	45	hsa-mir-454	34727290
21	hsa-mir-181c	31078780	46	hsa-mir-381	32518523
22	hsa-mir-212	33187481	47	hsa-mir-650	33086498
23	hsa-mir-330	dbDEMCM	48	hsa-mir-208a	30911286
24	hsa-mir-211	35296964	49	hsa-mir-33a	34950209
25	hsa-mir-372	30570852	50	hsa-mir-503	30594253

which we can see that target gene sets associated with hsa-mir-19b-2 were significantly enriched in several terms closely related to breast cancer, including regulation of hormone secretion, response to steroid hormone, insulin signaling and MAPK6/MAPK4 signaling. To be more specific, breast cancer growth is dependent upon estrogenic hormones and can be inhibited by antiestrogenic antagonists [64]. A lot of evidence also suggests that interactions between steroid hormones and growth factors act as a regulator of endocrine response in breast cancer, while abnormalities in growth factor signaling are common contributors to the endocrine-resistant phenotype [65]. Furthermore,

the dysregulation of the expression and function of insulin and its downstream signaling effectors often drives breast cancer initiation and progression in a subtype-dependent manner [66]. In addition, the high expression of MAPK4 defines a large subset or subtype of triple-negative breast cancer (TNBC) that responds to MAPK4 blockade, and targeting MAPK4 both inhibits TNBC growth and sensitizes tumors to PI3K blockade [67]. Finally, we observed that many biological processes and other terms were obtained based on the above enrichment analysis. Through utilizing Metascape, we further captured the relationships between these terms. As shown in Supplementary Figure S8, we selected a subset of enriched terms and presented them as a network graph, where terms with similarity  $> 0.3$  are connected by edges. According to the network graph, we discovered that several terms associated with breast cancer clustered together. In summary, the aforementioned biological analysis suggests a potential association between hsa-mir-19b-2 and the occurrence and progression of breast cancer.

## Conclusion

Identification of disease-related miRNAs through computational models can contribute to accelerating the understanding of the pathogenic mechanisms of human diseases and the discovery of potential therapeutic targets. In this study, we develop a SPLHRNMTF model of robust orthogonal non-negative matrix tri-factorization with self-paced learning and dual hypergraph regularization for predicting miRNA-disease associations. To be more specific, we first utilize a non-linear fusion method to obtain comprehensive similarity for miRNAs and diseases. Then, weighted  $k$ -nearest neighbor profile method is used to replace zero values in the association matrix with likelihood scores for correct false-negative associations. Furthermore, SPLHRNMTF makes use of  $L_{2,1}$  norm instead of Frobenius norm to calculate residual error for effectively alleviating the impact of noise and outliers on model prediction performance. Meanwhile, we integrate self-paced learning into NMTF using  $L_{2,1}$  norm to alleviate the model from falling into bad local optimal solutions. Finally, hypergraph regularization is introduced to preserve the high-order complex relations of biological similarity networks in the low-dimensional embedding space. The experimental results of 5-fold cross-validation five times and ablation studies show that SPLHRNMTF outperforms other baseline models and has good robustness. Each module in SPLHRNMTF and the preprocessing step are also demonstrated to be effective. Furthermore, the meaningful results of case studies



**Fig. 4** The enrichment analysis of target gene sets related to hsa-mir-19b-2

on breast neoplasms and lung neoplasms indicate that 48 and 46 of the top 50 predicted disease-related miRNAs are confirmed, further demonstrating that the SPLHRNMTF model can accurately infer miRNA-disease associations. In addition, unconfirmed miRNA-disease associations have biological significance through enrichment analysis. To sum up, SPLHRNMTF can be used as an effective model to predict the associations between miRNAs and diseases.

### Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s12864-024-10729-w>.

Supplementary Material 1

### Acknowledgements

Not applicable.

### Authors' contributions

DOY designed the framework, conducted the experiments, and wrote the manuscript. RM, JZ, XL, NA, PKW, JH, and JQZ modified the manuscript. All author(s) revised and approved the manuscript.

### Funding

This work was supported by Guangdong Medical Science and Technology Research Fund (No. A2024169), Guangdong Medical University Youth Scientific Research Cultivation Fund (No. GDMUQ2023008), Dongguan Science and Technology Commissioner Fund, China (No. 2DK23001).

### Availability of data and materials

The first dataset MDAv2.0\_data is obtained by preprocessing the HMDD v2.0 database. The second dataset MDAv3.2\_data1 is released in [34]. The third dataset MDAv3.2\_data2 is obtained by preprocessing the HMDD v3.2

database. HMDD v2.0 and HMDD v3.2 databases can be downloaded from <http://www.cuilab.cn/hmdd>. The directed acyclic graph (DAG) with disease hierarchical relationships obtained from the MeSH database (<https://www.nlm.nih.gov/mesh/>). The miRNA functional similarity can be obtained from the MISIM 2.0 database (<http://www.lirmed.com/misim/>). The miRNA sequences can be obtained from the miRBase database (<https://www.mirbase.org/>). The datasets and source codes used in this study are freely available at <https://github.com/Ouyang-Dong/SPLHRNMTF>.

### Data availability

The datasets and source codes used in this study are freely available at <https://github.com/Ouyang-Dong/SPLHRNMTF>.

### Declarations

#### Ethics approval and consent to participate

Not applicable.

#### Consent for publication

Not applicable.

#### Competing interests

The authors declare no competing interests.

Received: 21 March 2024 Accepted: 20 August 2024

Published online: 20 September 2024

### References

- Bartel DP. MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell*. 2004;116(2):281–97.
- Cheng AM, Byrom MW, Shelton J, Ford LP. Antisense inhibition of human miRNAs and indications for an involvement of miRNA in cell growth and apoptosis. *Nucleic Acids Res*. 2005;33(4):1290–7.
- Karp X, Ambros V. Encountering microRNAs in cell fate signaling. *Science*. 2005;310(5752):1288–9.

4. Ma L, Teruya-Feldstein J, Weinberg RA. Tumour invasion and metastasis initiated by microRNA-10b in breast cancer. *Nature*. 2007;449(7163):682–8.
5. Taganov KD, Boldin MP, Chang KJ, Baltimore D. NF- $\kappa$ B-dependent induction of microRNA miR-146, an inhibitor targeted to signaling proteins of innate immune responses. *Proc Natl Acad Sci*. 2006;103(33):12481–6.
6. Cho WC, Chow AS, Au JS. MIR-145 inhibits cell proliferation of human lung adenocarcinoma by targeting EGFR and NUDT1. *RNA Biol*. 2011;8(1):125–31.
7. Zhou W, Fong MY, Min Y, Somlo G, Liu L, Palomares MR, et al. Cancer-secreted miR-105 destroys vascular endothelial barriers to promote metastasis. *Cancer Cell*. 2014;25(4):501–15.
8. Sarkar S, Jun S, Rellick S, Quintana D, Cavendish J, Simpkins J. Expression of microRNA-34a in Alzheimer's disease brain targets genes linked to synaptic plasticity, energy metabolism, and resting state network activity. *Brain Res*. 2016;1646:139–51.
9. Chen X, Liu MX, Yan GY. RWRMDA: predicting novel human microRNA-disease associations. *Mol BioSyst*. 2012;8(10):2792–8.
10. Xuan P, Han K, Guo M, Guo Y, Li J, Ding J, et al. Prediction of microRNAs associated with human diseases based on weighted k most similar neighbors. *PLoS ONE*. 2013;8(8):e70204.
11. Xuan P, Han K, Guo Y, Li J, Li X, Zhong Y, et al. Prediction of potential disease-associated microRNAs based on random walk. *Bioinformatics*. 2015;31(11):1805–15.
12. Chen X, Yan CC, Zhang X, You ZH, Deng L, Liu Y, et al. WBSMDA: within and between score for miRNA-disease association prediction. *Sci Rep*. 2016;6(1):21106.
13. You ZH, Huang ZA, Zhu Z, Yan GY, Li ZW, Wen Z, et al. PBMDA: a novel and effective path-based computational model for miRNA-disease association prediction. *PLoS Comput Biol*. 2017;13(3):e1005455.
14. Chen M, Zhang Y, Li A, Li Z, Liu W, Chen Z. Bipartite heterogeneous network method based on co-neighbor for miRNA-disease association prediction. *Front Genet*. 2019;10:385.
15. Li A, Deng Y, Tan Y, Chen M. A novel miRNA-disease association prediction model using dual random walk with restart and space projection federated method. *PLoS ONE*. 2021;16(6):e0252971.
16. Chen X, Li TH, Zhao Y, Wang CC, Zhu CC. Deep-belief network for predicting potential miRNA-disease associations. *Brief Bioinform*. 2021;22(3):bbaa186.
17. Ha J, Park C. MLMD: Metric learning for predicting miRNA-disease associations. *IEEE Access*. 2021;9:78847–58.
18. Ha J, Park S. NCMD: Node2vec-based neural collaborative filtering for predicting miRNA-disease association. *IEEE/ACM Trans Comput Biol Bioinforma*. 2022;20(2):1257–68.
19. Li J, Zhang S, Liu T, Ning C, Zhang Z, Zhou W. Neural inductive matrix completion with graph convolutional networks for miRNA-disease association prediction. *Bioinformatics*. 2020;36(8):2538–46.
20. Tang X, Luo J, Shen C, Lai Z. Multi-view multichannel attention graph convolutional network for miRNA-disease association prediction. *Brief Bioinform*. 2021;22(6):bbab174.
21. Peng W, He Z, Dai W, Lan W. MHCLMDA: multihypergraph contrastive learning for miRNA-disease association prediction. *Brief Bioinform*. 2024;25(1):bbad524.
22. Xiao Q, Luo J, Liang C, Cai J, Ding P. A graph regularized non-negative matrix factorization method for identifying microRNA-disease associations. *Bioinformatics*. 2018;34(2):239–48.
23. Chen X, Wang L, Qu J, Guan NN, Li JQ. Predicting miRNA-disease association based on inductive matrix completion. *Bioinformatics*. 2018;34(24):4256–65.
24. Chen X, Yin J, Qu J, Huang L. MDHGI: matrix decomposition and heterogeneous graph inference for miRNA-disease association prediction. *PLoS Comput Biol*. 2018;14(8):e1006418.
25. Dissez G, Ceddia G, Pinoli P, Ceri S, Masseroli M. Drug repositioning predictions by non-negative matrix tri-factorization of integrated association data. In: *Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*. New York: ACM (Association for Computing Machinery); 2019. pp. 25–33.
26. Jamali AA, Kusaliik A, Wu F. NMTF-DTI: A nonnegative matrix tri-factorization approach with multiple kernel fusion for drug-target interaction prediction. *IEEE/ACM Trans Comput Biol Bioinforma*. 2021;20:586–94.
27. Zheng X, Zhang C, Wan C. MiRNA-Disease association prediction via non-negative matrix factorization based matrix completion. *Signal Proc*. 2022;190:108312.
28. Ha J, Park C, Park C, Park S. IMIPMF: Inferring miRNA-disease interactions using probabilistic matrix factorization. *J Biomed Inform*. 2020;102:103358.
29. Ha J. MDMF: predicting miRNA-disease association based on matrix factorization with disease similarity constraint. *J Personalized Med*. 2022;12(6):885.
30. Ha J. SMAP: Similarity-based matrix factorization framework for inferring miRNA-disease association. *Knowl-Based Syst*. 2023;263:110295.
31. Ha J. LncRNA expression profile-based matrix factorization for identifying lncRNA-disease associations. *IEEE Access*. 2024;12:70297–304.
32. Li Y, Qiu C, Tu J, Geng B, Yang J, Jiang T, et al. HMDD v2. 0: a database for experimentally supported human microRNA and disease associations. *Nucleic Acids Res*. 2014;42(D1):D1070–4.
33. Huang Z, Shi J, Gao Y, Cui C, Zhang S, Li J, et al. HMDD v3. 0: a database for experimentally supported human microRNA-disease associations. *Nucleic Acids Res*. 2019;47(D1):D1013–7.
34. Ning Q, Zhao Y, Gao J, Chen C, Li X, Li T, et al. AMHMDA: attention aware multi-view similarity networks and hypergraph learning for miRNA-disease associations identification. *Brief Bioinform*. 2023;24(2):bbad094.
35. Kozomara A, Birgaoanu M, Griffiths-Jones S. miRBase: from microRNA sequences to function. *Nucleic Acids Res*. 2019;47(D1):D1155–62.
36. Li J, Zhang S, Wan Y, Zhao Y, Shi J, Zhou Y, et al. MISIM v2. 0: a web server for inferring microRNA functional similarity based on microRNA-disease associations. *Nucleic Acids Res*. 2019;47(W1):W536–41.
37. 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):1644–50.
38. Zhao Y, Chen X, Yin J. Adaptive boosting-based computational model for predicting potential miRNA-disease associations. *Bioinformatics*. 2019;35(22):4730–8.
39. Liang C, Yu S, Luo J. Adaptive multi-view multi-label learning for identifying disease-associated candidate miRNAs. *PLoS Comput Biol*. 2019;15(4):e1006931.
40. Van Laarhoven T, Nabuurs SB, Marchiori E. Gaussian interaction profile kernels for predicting drug-target interaction. *Bioinformatics*. 2011;27(21):3036–43.
41. Chen X, Xie D, Wang L, Zhao Q, You ZH, Liu H. BNPMDA: bipartite network projection for miRNA-disease association prediction. *Bioinformatics*. 2018;34(18):3178–86.
42. Wang B, Mezzini AM, Demir F, Fiume M, Tu Z, Brudno M, et al. Similarity network fusion for aggregating data types on a genomic scale. *Nat Methods*. 2014;11(3):333–7.
43. Zhao Q, Meng D, Jiang L, Xie Q, Xu Z, Hauptmann A. Self-paced learning for matrix factorization. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Palo Alto: AAAI Press; vol. 29. 2015.
44. Meng D, Zhao Q, Jiang L. A theoretical understanding of self-paced learning. *Inf Sci*. 2017;414:319–28.
45. Huang S, Zhao P, Ren Y, Li T, Xu Z. Self-paced and soft-weighted non-negative matrix factorization for data representation. *Knowl-Based Syst*. 2019;164:29–37.
46. Cai D, He X, Han J, Huang TS. Graph regularized nonnegative matrix factorization for data representation. *IEEE Trans Pattern Anal Mach Intel*. 2010;33(8):1548–60.
47. Zeng K, Yu J, Li C, You J, Jin T. Image clustering by hyper-graph regularized non-negative matrix factorization. *Neurocomputing*. 2014;138:209–17.
48. Jiao CN, Gao YL, Yu N, Liu JX, Qi LY. Hyper-graph regularized constrained NMF for selecting differentially expressed genes and tumor classification. *IEEE J Biomed Health Inform*. 2020;24(10):3002–11.
49. Ezzat A, Zhao P, Wu M, Li XL, Kwok CK. Drug-target interaction prediction with graph regularized matrix factorization. *IEEE/ACM Trans Comput Biol Bioinforma*. 2016;14(3):646–56.
50. Wang MN, Xie XJ, You ZH, Ding DW, Wong L. A weighted non-negative matrix factorization approach to predict potential associations between drug and disease. *J Transl Med*. 2022;20(1):552.
51. Chen G, Wang F, Zhang C. Collaborative filtering using orthogonal non-negative matrix tri-factorization. *Inf Process Manag*. 2009;45(3):368–79.
52. Del Buono N, Pio G. Non-negative matrix tri-factorization for co-clustering: an analysis of the block matrix. *Inf Sci*. 2015;301:13–26.

53. Kumar M, Packer B, Koller D. Self-paced learning for latent variable models. *Adv Neural Inf Process Syst.* 2010;23:1189–97.
54. Jiang L, Meng D, Zhao Q, Shan S, Hauptmann A. Self-paced curriculum learning. In: *Proceedings of the AAAI Conference on Artificial Intelligence.* Palo Alto: AAAI Press; vol. 29. 2015.
55. Ding CH, Li T, Jordan MI. Convex and semi-nonnegative matrix factorizations. *IEEE Trans Pattern Anal Mach Intel.* 2008;32(1):45–55.
56. Ding C, Li T, Peng W, Park H. Orthogonal nonnegative matrix t-factorizations for clustering. In: *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining.* New York: ACM (Association for Computing Machinery); 2006. pp. 126–135.
57. Kong D, Ding C, Huang H. Robust nonnegative matrix factorization using l21-norm. In: *Proceedings of the 20th ACM international conference on Information and knowledge management.* New York: ACM (Association for Computing Machinery); 2011. pp. 673–682.
58. Dang Q, Liang Y, Ouyang D, Miao R, Ling C, Liu X, et al. Improved Computational Drug-Repositioning by Self-Paced Non-Negative Matrix Tri-Factorization. *IEEE/ACM Trans Comput Biol Bioinforma.* 2022;20:1953–62.
59. Cui C, Zhong B, Fan R, Cui Q. HMDD v4. 0: a database for experimentally supported human microRNA-disease associations. *Nucleic Acids Res.* 2024;52(D1):D1327–32.
60. Xu F, Wang Y, Ling Y, Zhou C, Wang H, Teschendorff AE, et al. dbDEM3.0: functional exploration of differentially expressed miRNAs in cancers of human and model organisms. *Genomics Proteomics Bioinforma.* 2022;20(3):446–54.
61. Bockhorn J, Prat A, Chang YF, Liu X, Huang S, Shang M, et al. Differentiation and loss of malignant character of spontaneous pulmonary metastases in patient-derived breast cancer models. *Cancer Res.* 2014;74(24):7406–17.
62. Huang HY, Lin YCD, Cui S, Huang Y, Tang Y, Xu J, et al. miRTarBase update 2022: an informative resource for experimentally validated miRNA-target interactions. *Nucleic Acids Res.* 2022;50(D1):D222–30.
63. Zhou Y, Zhou B, Pache L, Chang M, Khodabakhshi AH, Tanaseichuk O, et al. Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. *Nat Commun.* 2019;10(1):1523.
64. Dickson RB, Lippman ME. Estrogenic regulation of growth and polypeptide growth factor secretion in human breast carcinoma. *Endocr Rev.* 1987;8(1):29–43.
65. Nicholson RI, McClelland RA, Robertson J, Gee J. Involvement of steroid hormone and growth factor cross-talk in endocrine response in breast cancer. *Endocr-Relat Cancer.* 1999;6(3):373–87.
66. Lero MW, Shaw LM. Diversity of insulin and IGF signaling in breast cancer: implications for therapy. *Mol Cell Endocrinol.* 2021;527:111213.
67. Wang W, Han D, Cai Q, Shen T, Dong B, Lewis MT, et al. MAPK4 promotes triple negative breast cancer growth and reduces tumor sensitivity to PI3K blockade. *Nat Commun.* 2022;13(1):245.

## Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.