frontiers in Genetics

ORIGINAL RESEARCH
published: 14 October 2021
doi: 10.3389/fgene.2021.754425

Multi-Similarities Bilinear Matrix Factorization-Based Method for Predicting Human Microbe–Disease Associations

Xiaoyu Yang[1,2], Linai Kuang[1]*, Zhiping Chen[2] and Lei Wang[1,2]*

[1]Key Laboratory of Hunan Province for Internet of Things and Information Security, Xiangtan University, Xiangtan, China, [2]College of Computer Engineering and Applied Mathematics, Changsha University, Changsha, China

Accumulating studies have shown that microbes are closely related to human diseases. In this paper, a novel method called MSBMFHMDA was designed to predict potential microbe–disease associations by adopting multi-similarities bilinear matrix factorization. In MSBMFHMDA, a microbe multiple similarities matrix was constructed first based on the Gaussian interaction profile kernel similarity and cosine similarity for microbes. Then, we use the Gaussian interaction profile kernel similarity, cosine similarity, and symptom similarity for diseases to compose the disease multiple similarities matrix. Finally, we integrate these two similarity matrices and the microbe-disease association matrix into our model to predict potential associations. The results indicate that our method can achieve reliable AUCs of 0.9186 and 0.9043 ± 0.0048 in the framework of leave-one-out cross validation (LOOCV) and fivefold cross validation, respectively. What is more, experimental results indicated that there are 10, 10, and 8 out of the top 10 related microbes for asthma, inflammatory bowel disease, and type 2 diabetes mellitus, respectively, which were confirmed by experiments and literatures. Therefore, our model has favorable performance in predicting potential microbe–disease associations.

Keywords: microbe, disease, association prediction, multi-similarities, matrix factorization

OPEN ACCESS

Edited by:
Juan Caballero,
European Bioinformatics Institute (EMBL-EBI), United Kingdom

Reviewed by:
Min Chen,
Hunan Institute of Technology, China
Yushan Qiu,
Shenzhen University, China

***Correspondence:**
Linai Kuang
kla@xtu.edu.cn
Lei Wang
wanglei@xtu.edu.cn

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

Received: *06 August 2021*
Accepted: *21 September 2021*
Published: *14 October 2021*

Citation:
Yang X, Kuang L, Chen Z and Wang L (2021) Multi-Similarities Bilinear Matrix Factorization-Based Method for Predicting Human Microbe–Disease Associations. Front. Genet. 12:754425. doi: 10.3389/fgene.2021.754425

INTRODUCTION

Microorganisms are the general names of all tiny organisms that individuals cannot observe with the naked eye, but are closely related to humans. Microorganisms include bacteria, viruses, fungi, and a large group of small protozoa, microalgae (The Human Microbiome Project Consortium, 2012). We all know that microbes can cause diseases and make food, cloth, and leather moldy and decay, but it also has a beneficial side. For instance, probiotics in the gut are beneficial to ferment undigested carbohydrates in order to produce nutrition needed for the human body. One of the most important effects of microbes on human beings is to lead to the spread of infectious diseases. Viruses are the cause of 50% of human diseases, therefore, microbes can greatly influence human health. For example, *Mycobacterium tuberculosis* and *Bacillus anthracis* can cause tuberculosis and anthrax, respectively (Hawn et al., 2014; Hendricks et al., 2014). Therefore, identifying disease-related microbes is one of the important tasks in the study of complex disease pathology. One of the useful values of biological research is its application in the field of medicine for the benefit of human health. Identification and prediction of human microbe–disease associations are important for disease prevention, diagnosis, treatment, and prognosis. Nevertheless, the traditional test methods are time consuming and costly. As the result, it is crucial to predict microbe–disease associations by computational methods.

Due to the rapid development of artificial intelligence (AI) and machine learning technology (Huang, 1996; Huang, 1999; Huang and Du, 2008), many computational methods are widely applied in predicting the potential correlation among biological entities [such as miRNA-disease (Chen and Yan, 2015; You et al., 2017; Chen et al., 2018a; Chen et al., 2018b), lncRNA-disease (Chen and Yan, 2013; Chen et al., 2016b; Yu et al., 2018; Chen et al., 2019; Xuan et al., 2019), and drug–target interaction prediction (Chen et al., 2012)]. Meanwhile, many computational methods have been proposed to predict microbe–disease associations. According to the introduction of this paper (Wen et al., 2021), the existing methods can be divided into five categories, namely, path-based methods, random walk methods, bipartite local models, matrix factorization methods, and other methods. The path-based method mainly calculates the relationship between microbe and disease by two indexes, one is walk length, the other is the number of paths reached. KATZHMDA (Chen et al., 2016a), based on path-based method, is the first calculation method by computing the number of walks of connections between microbe and disease nodes in the microbe–disease association network. Random walk methods first construct a transition probability network by microbe and disease nodes; a potential association is then searched by measuring the path probability of the walker from the start node to the end node in the network. BiRWHMDA (Zou et al., 2017), BiRWMP (Shen et al., 2018), and NBLPIHMDA (Wang et al., 2019) using random walk achieves satisfying performance. Bipartite local models calculate the prediction scores of microbes and diseases, respectively, and then the two scores are combined as the final prediction score. Matrix factorization methods decompose an interaction matrix into two low dimensional matrices representing disease features and microbe features. Finally, the product of the two feature matrices is taken as the final prediction matrix. CMFHMDA (Shen et al., 2017) is the first calculation model based on matrix factorization by integrating known microbe–disease association and Gaussian interaction profile kernel similarity for microbes and diseases. MDLPHMDA (Qu et al., 2019) puts forward the matrix decomposition and label propagation to predict microbe–disease association. NMFMDA (Liu et al., 2018) predicts potential associations by graph-regularized non-negative matrix factorization. Other methods mainly include ensemble learning and matrix completion, such as ABHMDA (Peng et al., 2018), BMCMDA (Shi et al., 2018), and MCHMDA (Yan et al., 2021). What is more, the methods based on matrix decomposition were developed to predict the relationship between other biological entities (Wang and Gao, 2015; Qiu et al., 2021a; Qiu et al., 2021b), for example, Qiu et al. (2021a) proposed a novel model based on weighted data fusion with sparse matrix tri-factorization to predict associations between RNA-binding proteins and alternative splicing, namely, WDFSMF. WDFSMF simultaneously decomposes heterogeneous data source matrices into low-rank matrices to mine potential associations.

However, some of the above prediction models of microbe–disease have their own limitations. Owing to the lack of measurements for microbe and disease similarity, some models, which are only based on the Gaussian interaction profile kernel similarity of microbes and diseases, cannot be used to predict diseases that are not associated with microbes. In this study, considering the above limitations and inspired by the good performance of multi-similarities bilinear matrix factorization method to predict drug-associated indications (Yang et al., 2021), we proposed a new microbe–disease association prediction model called MSBMFHMDA. The overall workflow of our method is illustrated in **Figure 1**. First, we calculated the Gaussian interaction profile kernel similarity and cosine similarity for diseases and microbes based on the dataset of known microbe–disease associations. Then, two concatenated microbe and disease similarity matrices are constructed based on the Gaussian interaction profile kernel similarity for diseases and microbes, disease symptom similarity, cosine similarity for diseases, and microbes. Notably, we concatenate these similarity matrices of microbe and disease instead of fusing multiple similarities into a single similarity matrix. Finally, we integrate these two concatenated similarity matrices and the microbe–disease association matrix into our MSBMF model to infer potential microbe–disease associations. The framework of LOOCV and fivefold cross validation were implemented to estimate the prediction performances of MSBMFHMDA. The results suggested that our method could achieve reliable AUCs of 0.9186 and 0.9043 ± 0.0048 in LOOCV and fivefold cross validation, respectively, which is much better than state-of-the-art methods. Moreover, we further implemented the case studies of asthma, IBD, and T2D on MSBMFHMDA, and the reliability of our model is further verified.

MATERIALS AND METHODS

Datasets

The Human Microbe–Disease Association Database (HMDAD) (Ma et al., 2017) is the first human microbe–disease association database established by Ma et al. through a lot of biological experiments. The database includes 483 experimentally tested and verified associations between 292 microbes and 39 diseases. We downloaded the data from HMDAD (http://www.cuilab.cn/hmdad), then removed redundant associations. Thus, 450 microbe–disease associations including 39 diseases and 292 microbes were obtained from 61 publications. As a result, a 39 × 292 dimensional adjacency matrix A is constructed. In addition, in the adjacency matrix A, the value of $A[i][j]$ is set to 1 if microbe $m[j]$ is related to disease $d[i]$, otherwise, $A[i][j]$ is set to 0.

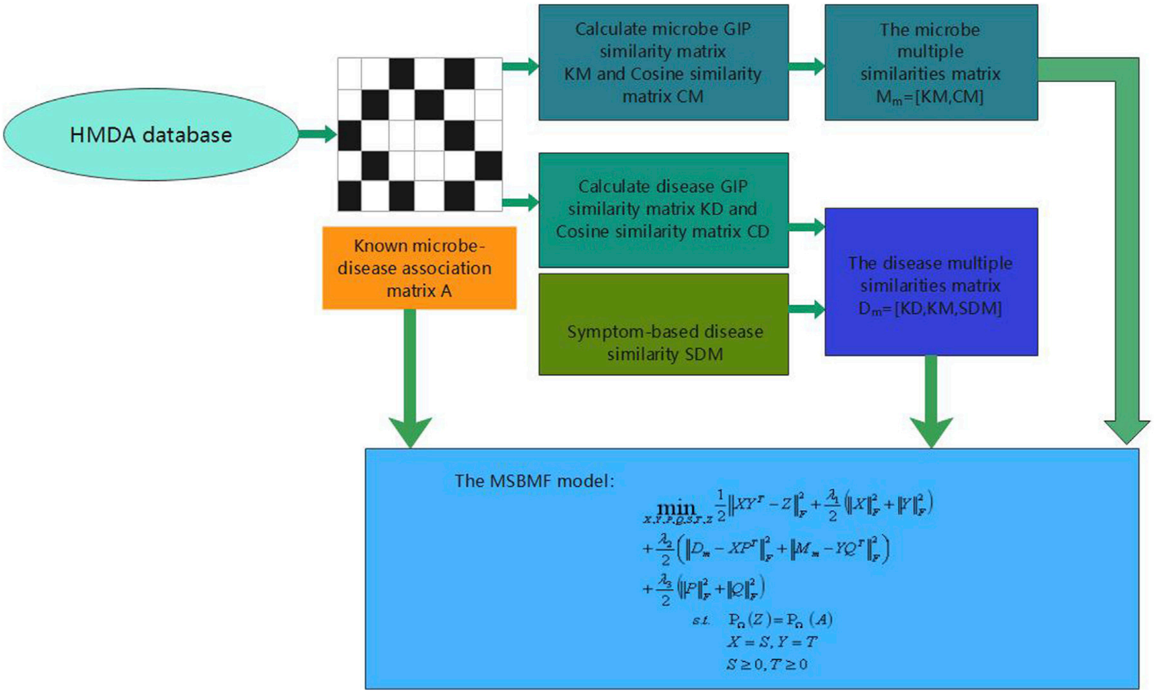
HMDA database
Calculate microbe GIP similarity matrix KM and Cosine similarity matrix CM
The microbe multiple similarities matrix M_m=[KM,CM]
Calculate disease GIP similarity matrix KD and Cosine similarity matrix CD
Known microbe-disease association matrix A
The disease multiple similarities matrix D_m=[KD,KM,SDM]
Symptom-based disease similarity SDM
The MSBMF model:
$\min_{X,Y,P,Q,S,T,Z} \frac{1}{2}\|XY^T - Z\|_F^2 + \frac{\lambda_1}{2}(\|X\|_F^2 + \|Y\|_F^2)$
$+\frac{\lambda_2}{2}(\|D_m - XP^T\|_F^2 + \|M_m - YQ^T\|_F^2)$
$+\frac{\lambda_3}{2}(\|P\|_F^2 + \|Q\|_F^2)$
$s.t.\quad P_\Omega(Z) = P_\Omega(A)$
$X = S, Y = T$
$S \geq 0, T \geq 0$

FIGURE 1 | The overall workflow of MSBMFHMDA.

Similarity Measures of Microbe

Gaussian Interaction Profile Kernel Similarity of Microbes: *KM*

Gaussian kernel function is a common kernel function. Its essence is to measure the similarity between samples (van Laarhoven et al., 2011). It is based on the assumption that two similar diseases and the same microbe will exhibit the same interaction and non-interaction relationship. Therefore, in the known microbe–disease association network, we adopt the Gaussian interaction profile kernel similarity to compute microbe similarity according to the following **Eq. 1**:

$$\mathrm{KM}(m(i), m(j)) = \exp(-\gamma_m \|IP(m(i) - IP(m(j))\|^2) \tag{1}$$

where $m[i]$ and $m[j]$ represent the *ith* and *jth* microbes, respectively, in the matrix A, and its interaction profiles $IP(m(i))$ and $IP(m(j))$ represent the *ith* and *jth* column, respectively. Based on this information, we can calculate the similarity between the two microbe vectors by calculating the *L2* norm. Additionally, the parameter γ_m can be calculated as follows:

$$\gamma_\mathrm{m} = \frac{\gamma_m'}{\left(\frac{1}{n_m}\sum_{k=1}^{n_m}\|IP(m(k))\|^2\right)} \tag{2}$$

where γ_m is a parameter used to control the bandwidth of the Gaussian kernel function; it is the result of normalization by bandwidth parameter γ_m', and according to the previous experiment (van Laarhoven et al., 2011), γ_m' will be set to 1. n_m is the total number of microbes collected from the HMDAD, so, n_m is equal to 292.

Cosine Similarity of Microbes: *CM*

Microbe cosine similarity is calculated based on assumptions that if the microbes are similar to each other (Xie et al., 2019). In other words, in the microbe–disease association matrix, $A(i,:)$ and $A(j,:)$ should be similar to each other. Therefore, the cosine similarity between microbe $m(i)$ and microbe $m(j)$ can be calculated as follows:

$$CM(m(i), m(j)) = \frac{A(i,:)\cdot A(j,:)}{\|A(i,:)\| \times \|A(j,:)\|} \tag{3}$$

where $A(i,:)$ represents the *ith* row of adjacency matrix A; the result is then projected into [0, 1] by the min–max normalization.

Similarity Measures of Disease

Gaussian Interaction Profile Kernel Similarity of Diseases: *KD*

In a similar way, the Gaussian interaction profile kernel similarity between disease $\mathrm{d}(i)$ and disease $\mathrm{d}(j)$ can be defined as follows:

$$\mathrm{KD}(\mathrm{d}(i), \mathrm{d}(j)) = \exp(-\gamma_\mathrm{d}\|IP(\mathrm{d}(i) - IP(\mathrm{d}(j))\|^2) \tag{4}$$

$$\gamma_\mathrm{d} = \frac{\gamma_\mathrm{d}'}{\left(\frac{1}{n_\mathrm{d}}\sum_{k=1}^{n_\mathrm{d}}\|IP(\mathrm{d}(k))\|^2\right)} \tag{5}$$

γ_d' will be also set to 1; $\mathrm{n_d}$ is equal to 39.

Cosine Similarity of Diseases: *CD*

The cosine similarity between disease $\mathrm{d}(i)$ and disease $\mathrm{d}(j)$ is given as follows:

$$CD\left(\mathrm{d}(i), \mathrm{d}(j)\right) = \frac{A(:,\mathrm{i}) \cdot A(:,\mathrm{j})}{\|A(:,\mathrm{i})\| \times \|A(:,\mathrm{j})\|} \tag{6}$$

where $A(:, i)$ represents the ith column of adjacency matrix A; the result is then projected into [0, 1] by the min–max normalization.

Symptom-Based Disease Similarity: *SDM*

The abnormal subjective feeling or some objective pathological changes of patients caused by a series of abnormal changes in function, metabolism, and morphological structure in the process of disease are called symptoms. Some diseases, especially in the early stage of some diseases, may not be accompanied by symptoms and signs. The human symptoms–disease network (HSDN) has been constructed by Zhou et al. from PubMed (Wheeler et al., 2007; Zhou et al., 2014). Moreover, they used term frequency inverse document frequency (TF-IDF) (Salton et al., 1975) to measure the symptom-based disease similarity based on the co-occurrence frequency between a disease and a symptom. Based on these data, Chen et al. (2016a) extracted those symptom-based similarities of common diseases from HMDAD. Hence, symptom similarity SDM can be constructed.

MSBMF Model

As the microbe–disease association matrix is low rank, in other words, it is very sparse, microbe-disease association matrix can be split into two low-dimensional feature matrices, i.e., disease feature X and microbe Y. Then, Tikhonov regularization terms are used to avoid over-fitting. The elementary matrix factorization model is formulated as follows:

$$\min_{X,Y} \frac{1}{2}\left\|P_\Omega\left(XY^T - A\right)\right\|_F^2 + \frac{\lambda_1}{2}\left(\|X\|_F^2 + \|Y\|_F^2\right) \tag{7}$$

where $\| \bullet \|_F$denotes the Frobenius norm, $\|A\|_F = \sqrt{tr\left(A^T A\right)} = \sqrt{\sum_{i=1}^m \sum_{j=1}^n a_{ij}^2}$, tr(A) is the trace of matrix A, $\|A\|_F^2 = tr\left(A^T A\right) = \sum_{i=1}^m \sum_{j=1}^n a_{ij}^2$, λ_1 is the harmonic parameter to counterpoise the error term and the regularization terms, Ω is an index set of known association in matrix A, and P_Ω is defined as:

$$\left(P_\Omega(I)\right)_{ij} = \begin{cases} I_{ij}, & (i,j) \in \Omega \\ 0, & (i,j) \notin \Omega \end{cases} \tag{8}$$

However, **Eq. 7** does not involve prior information about diseases and microbes. Given a disease similarity matrix D and a microbe similarity matrix M, as X,Y can be considered as matrices containing disease and microbe potential characteristic vectors, respectively, XX^T and YY^T are expected to match D and M, respectively (Zheng et al., 2013; Cui et al., 2019). Therefore, **Eq. 7** is extended to:

$$\min_{X,Y} \frac{1}{2}\left\|P_\Omega\left(XY^T - A\right)\right\|_F^2 + \frac{\lambda_1}{2}\left(\|X\|_F^2 + \|Y\|_F^2\right) + \frac{\lambda_2}{2}\left(\left\|D - XX^T\right\|_F^2 + \left\|M - YY^T\right\|_F^2\right) \tag{9}$$

TABLE 1 | The area under the curve (AUC) value using different λ_1 and λ_2 values in the leave-one-out cross validation (LOOCV).

λ_2 / λ_1	0.001	0.01	0.1	1
0.001	0.8667	0.8653	0.7689	0.6894
0.01	0.8849	0.8884	0.8798	0.7854
0.1	0.9067	0.9186	0.8968	0.8764
1	0.8932	0.8946	0.8937	0.8831

TABLE 2 | The AUC value using different τ values while fixing λ_1 = 0.1 and λ_2 = 0.01.

τ	0.1	0.3	0.5	0.7	0.9	1
AUC	0.8556	0.8721	0.8901	0.9186	0.9187	0.9186

In order to incorporate multiple similarity measures, an MSBMF model can be proposed for predicting microbe–disease associations, which is formulated as follows:

$$\begin{aligned} \min_{X,Y,P,Q,Z} \; & \frac{1}{2}\left\|XY^T - Z\right\|_F^2 + \frac{\lambda_1}{2}\left(\|X\|_F^2 + \|Y\|_F^2\right) \\ & + \frac{\lambda_2}{2}\left(\left\|D_m - XP^T\right\|_F^2 + \left\|M_m - YQ^T\right\|_F^2\right) \\ & + \frac{\lambda_3}{2}\left(\|P\|_F^2 + \|Q\|_F^2\right) \\ s.t. \quad & P_\Omega(A) = P_\Omega(Z) \\ & X \geq 0, Y \geq 0 \end{aligned} \tag{10}$$

D_m and M_m are multi-similarities matrices of diseases and microbes, respectively, and λ_1, λ_2, λ_3 are balancing parameters. Obviously, $D_m = [KD, CD, SDM]$ and $\mathrm{M_m} = [KM, CM]$, where P and Q are matrices including latent features representing disease similarity and microbe similarity, respectively. Z is an auxiliary matrix that helps to optimize. Furthermore, by introducing two splitting matrices S and T, **Eq. 10** is transformed into:

$$\begin{aligned} \min_{X,Y,P,Q,S,T,Z} \; & \frac{1}{2}\left\|XY^T - Z\right\|_F^2 + \frac{\lambda_1}{2}\left(\|X\|_F^2 + \|Y\|_F^2\right) \\ & + \frac{\lambda_2}{2}\left(\left\|D_m - XP^T\right\|_F^2 + \left\|M_m - YQ^T\right\|_F^2\right) \\ & + \frac{\lambda_3}{2}\left(\|P\|_F^2 + \|Q\|_F^2\right) \\ s.t. \quad & P_\Omega(A) = P_\Omega(Z) \\ & S = X, T = Y \\ & S \geq 0, T \geq 0 \end{aligned} \tag{11}$$

Then, we use the alternating direction method of multipliers (ADMM) framework to solve **Eq. 10**. The augmented Lagrangian function is given by:

$$\begin{aligned} \ell(\mathrm{X}, Y, P, Q, S, T, Z) = & \frac{1}{2}\left\|XY^T - Z\right\|_F^2 + \frac{\lambda_1}{2}\left(\|X\|_F^2 + \|Y\|_F^2\right) \\ & + \frac{\lambda_2}{2}\left(\left\|D_m - XP^T\right\|_F^2 + \left\|M_m - YQ^T\right\|_F^2\right) + \frac{\lambda_3}{2}\left(\|P\|_F^2 + \|Q\|_F^2\right) \\ & + \langle \Phi, X - S \rangle + \langle \Psi, Y - T \rangle + \frac{\mu}{2}\left(\|X - S\|_F^2 + \|Y - T\|_F^2\right) \end{aligned} \tag{12}$$

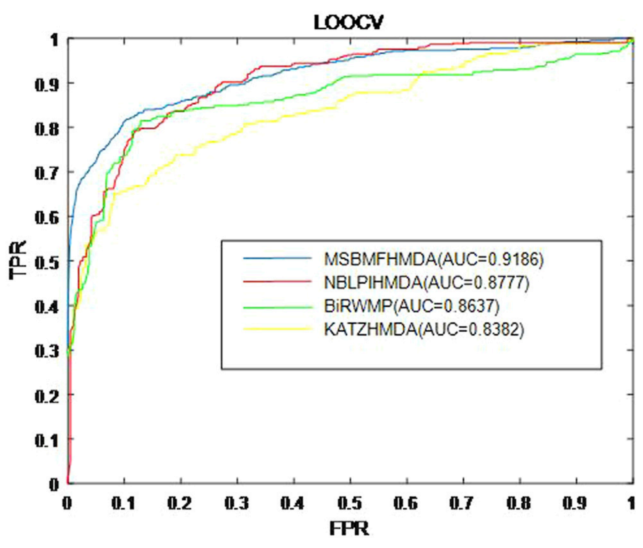

FIGURE 2 | Prediction performance comparison between MSBMFHMDA and the other three methods in leave-one-out cross validation (LOOCV).

TABLE 3 | Performances of different methods in LOOCV and fivefold CV.

Method	LOOCV	Five-fold CV
MSBMFHMDA	0.9186	0.8993 ± 0.0032
NBLPIHMDA	0.8777	0.8958 ± 0.0027
BiRWMP	0.8637	0.8522 ± 0.0054
KATZHMDA	0.8382	0.8301 ± 0.0033

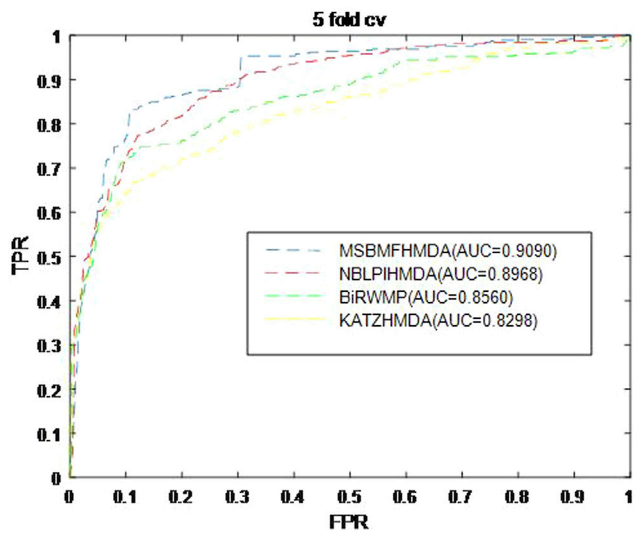

FIGURE 3 | Prediction performance comparison between MSBMFHMDA and the other three methods in fivefold cross validation.

where Φ and Ψ are the Lagrange multipliers, and μ is the penalty parameter. After k iteration, $X_{k+1}, Y_{k+1}, P_{k+1}, Q_{k+1,} S_{k+1}, T_{k+1}$ *and* Z_{k+1} will be computed. We adopt a scheme with gradually increasing learning rate to achieve fast convergence (Shang et al., 2018). After executing the MSBMF algorithm, a non-negative matrix M* is a predicted scores matrix. The scheme of MSBMF model is illustrated in **Algorithm 1**.

Algorithm 1. MSBMF algorithm.

```
Input: the microbe–disease association matrix M, the multiply similarities of disease matrices $D_m$, the multiply similarities of microbe matrices $M_m$, subspace dimensionality r, parameters $\lambda_1$, $\lambda_2$ and $\lambda_3$.
Output: predicted association matrix M*.
Step1: calculate microbe GIP similarity and cosine similarity;
Step2: calculate disease GIP similarity, cosine similarity, and symptom-based similarity;
Step 3: initializing randomly four non-negative matrices $X_0$, $Y_0$, $P_0$, $Q_0$; $S_0 = X_0$, $T_0 = Y_0$, $Z_0 = M$, $\Phi_0 = 0$, $\Psi_0 = 0$, $\mu_0$, $\mu_{max}$, and rate changing factor $\rho > 1$;
Step4: repeat compute $X_{k+1}$, $Y_{k+1}$, $P_{k+1}$, $Q_{k+1}$, $S_{k+1}$, $T_{k+1}$, and $Z_{k+1,}$ and update the multipliers by: $\Phi_{k+1} \leftarrow \Phi_k + \mu_k (X_{k+1} - S_{k+1})$; $\Psi_{k+1} \leftarrow \Psi_k + \mu_k (Y_{k+1} - T_{k+1})$; update $\mu_{k+1}$ by $\mu_{k+1} \leftarrow \min(\rho\mu_k, \mu_{max})$; $k \leftarrow k + 1$; until convergence;
Step5: obtain the predicted association matrix M*.
Step6: Return M*.
```

RESULTS

Performance Evaluation

The problem of microbe–disease associations prediction can be seen as a classification or regression problem, usually using cross-validation to evaluate the generalization capabilities of the new sample. In order to evaluate performance of our model, we carry out two kinds of computational experiments, including LOOCV and fivefold cross validation. In LOOCV, each confirmed microbe–disease association was chosen as a test sample in turn, and the rest of the associations were used to train. After executing MSBMFHMDA, the score of the test example would be ranked with the scores of candidate samples that were made up of all unconfirmed microbe–disease pairs. In fivefold cross validation, we first divided the known microbe–disease associations into five equal parts and later made each part as a test sample in turn and the remaining four parts of associations as training samples. Similarly, the score of each test sample would be ranked with the scores of candidate samples that were made up of all unconfirmed microbe–disease pairs. As the sample divisions may cause bias, we repeated the fivefold cross-validation 100 times to get an average value as the final result. As the predicted score that obtained a higher rank than the given threshold, our model is considered to make a successful prediction. Then according to diverse thresholds, we plotted the receiver operating characteristics (ROC) curve by computing the ratio of true positive rate (TPR, sensitivity) to false positive rate (FPR, 1-specificity). The AUC can be used to

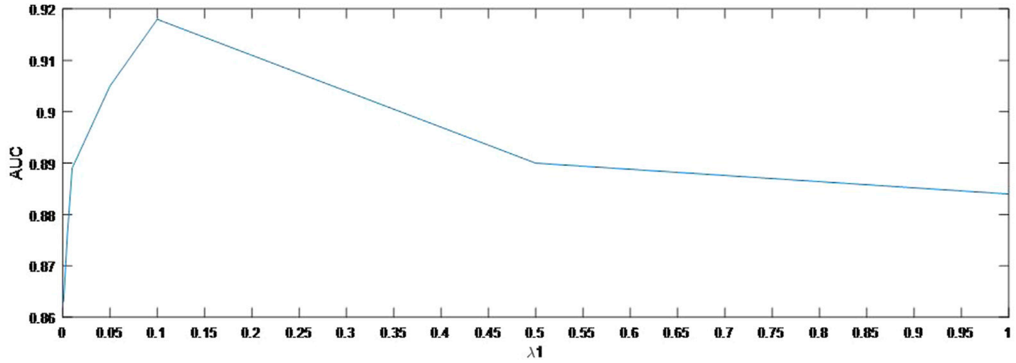

FIGURE 4 | Variation of the AUCs with the various settings of λ_1.

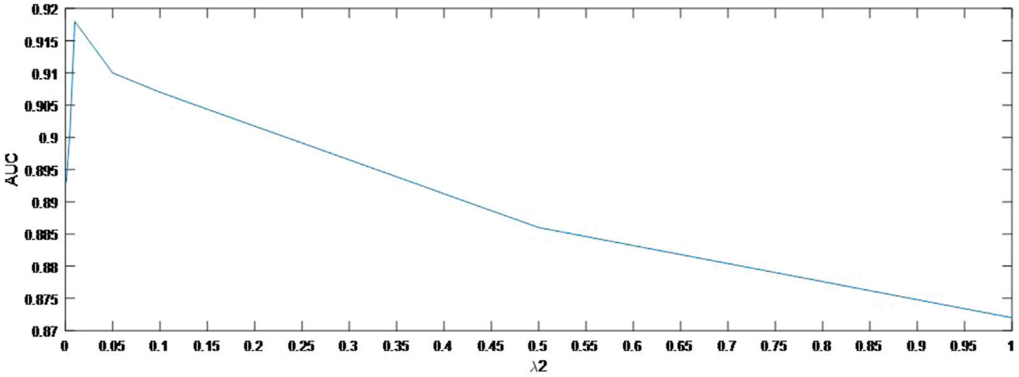

FIGURE 5 | Variation of the AUCs with the various settings of λ_2.

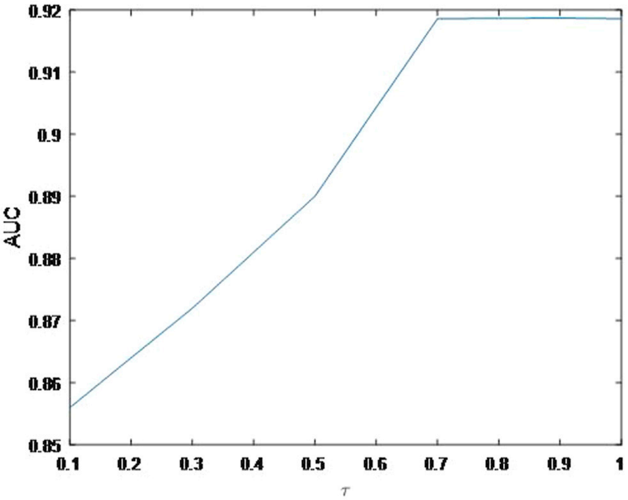

FIGURE 6 | Variation of the AUCs with the various settings of τ.

evaluate its predictive performance, where the AUC value of 1 represents perfect prediction ability, and the AUC value of 0.5 indicates random prediction performance (Chen et al., 2016a).

TABLE 4 | The validation results of the top 10 predicted asthma-related microbes by implementing MSBMFHMDA.

Rank	Microbe	Evidence
1	*Firmicutes*	PMID:23265859
2	*Clostridium difficile*	PMID:21872915
3	*Staphylococcus aureus*	PMID:17950502
4	*Bacteroides*	PMID:18822123
5	*Clostridium coccoides*	PMID:21477358
6	*Lachnospiraceae*	PMID:27433177
7	*Tropheryma whipplei*	PMID:26647445
8	*Lactobacillus*	PMID:20592920
9	*Actinobacteria*	PMID:23265859
10	*Enterobacteriaceae*	PMID:21639872

Effects of the Parameters

In our algorithm, the tunable parameters include the latent dimension r and the three coefficients λ_1, λ_2, and λ_3. We set r = [τmin (m, n)], where $\tau \in [0, 1]$ and [•] denotes the rounding function. Because there are many parameters, they may lead to overfitting. So, we set λ_2 and λ_3 to the same value to prevent overfitting. Finally, three parameters need to be determined, including τ, λ_1, and λ_2.

TABLE 5 | The validation results of the top 10 predicted inflammatory bowel disease (IBD)-related microbes by implementing MSBMFHMDA.

Rank	Microbe	Evidence
1	*Clostridium coccoides*	PMID:21477358
2	*Prevotella*	PMID:24013298
3	*Lactobacillus*	PMID:20592920
4	*Bacteroidetes*	PMID:29492876
5	*Veillonella*	PMID:30573380
6	*Clostridium difficile*	PMID:21872915
7	*Firmicutes*	PMID:23265859
8	*Staphylococcus aureus*	PMID:17950502
9	*Helicobacter pylori*	PMID:22221289
10	*Actinobacteria*	PMID:23265859

TABLE 6 | The validation results of the top 10 predicted type 2 diabetes (T2D)-related microbes by implementing MSBMFHMDA.

Rank	Microbe	Evidence
1	*Clostridium difficile*	PMID:21872915
2	*Enterobacteriaceae*	PMID:21639872
3	*Staphylococcus aureus*	PMID:17950502
4	*Helicobacter pylori*	PMID:22221289
5	*Prevotella*	PMID:24013298
6	*Veillonella*	Unconfirmed
7	*Lachnospiraceae*	PMID:27433177
8	*Bacteroides*	PMID:18822123
9	*Burkholderia*	Unconfirmed
10	*Actinobacteria*	PMID:23265859

We choose to adopt a "fixing one and determining the others" strategy. First, we set τ to 0.1 and then picked the values of λ_1 and λ_2 from {0.001, 0.01, 0.1, 1} by LOOCV in a standard dataset. Then, we fix the determined values of λ_1 and λ_2, and selected τ from {0.1,0.3,0.5,0.7,0.9,1}. The computational results for determining the λ_1 and λ_2 are listed in **Table 1**. We can discover that the AUC value reach maximum when $\lambda_1 = 0.1$ and $\lambda_2 = 0.01$. As shown in **Table 2**, our model furnishes approximately the same good performance when $\tau \geq 0.7$. Therefore, we set $\tau = 0.7$.

The stopping criteria of the MSBMF algorithm are $f_k \leq tol_1$ and $\frac{|f_{k+1}-f_k|}{\max\{1,|f_k|\} \leq tol_2}$, where $f_k = \frac{\|S_{k+1}T_{k+1}-S_kT_k\|_F}{\|S_kT_k\|_F}$ and tol_1, tol_2 are the given tolerances. Here, according to the related studies (Yang et al., 2021), we set $tol_1 = 2 \times 10^{-3}$ and $tol_2 = 10^{-4}$.

Comparison With Other State-of-the-Art Methods

In this section, we consider several state-of-the-art microbe–disease association prediction methods and make comparisons to demonstrate superior performance of our proposed method MSBMFHMDA. We compare it with KATZHMDA, BiRWMP, and NBLPIHMDA based on the dataset of known microbe–disease associations. As illustrated in the following **Figure 2** and **Table 3**, MSBMFHMDA yields best performance in LOOCV, achieving an AUC score of 0.9186, while KATZHMDA, BiRWMP, and NBLPIHMDA produce AUC scores of 0.8382, 0.8637, and 0.8777, respectively. As demonstrated in the following **Figure 3**, in the framework of fivefold cross validation, MSBMFHMDA can achieve a reliable AUC of 0.9043 ± 0.0048, which is better than the AUC achieved by KATZHMDA (0.8301 ± 0.0033), BiRWMP (0.8522 ± 0.0054), and NBLPIHMDA (0.8958 ± 0.0027).

The Sensitivity Analysis of Parameters

In this section, we concentrate on the sensitivity analysis for λ_1, λ_2, and τ in LOOCV. As we all know, when $\lambda_1 = 0.1, \lambda_2 = 0.01$, and $\tau = 0.7$, our model can realize excellent performance. We vary one parameter and keep the rest of the two parameters fixed to observe how the parameter benefits the AUC value.

As shown in **Figure 4**, the AUC can achieve the best values when $\lambda_1 = 0.1$. In the same way, **Figure 5** indicates the best AUC on $\lambda_2 = 0.01$. Finally, the effect of parameter τ on the prediction accuracy is discussed. **Figure 6** shows the AUC values of MSBMF with different τ. When $\tau > 0.7$, the trend of AUC is becoming steady. If τ continue to increase to 0.9 or 1, our model will not only generate overfitting but also increases the computational complexity.

Case Studies

Microbes are closely related to human health, and it is meaningful to explore whether microbes are associated with disease. In order to investigate into disease-causing microbes and further measure the prediction performance of our model, we selected three kinds of common microbe-induced diseases as cases for the analysis, namely, asthma, inflammatory bowel disease, and type 1 diabetes. The scores of the top 10 disease-related microbes are published in **Supplementary Tables S1–S3**, respectively.

Asthma is short for bronchial asthma, a heterogeneous disease characterized by chronic airway inflammation and airway hyper-responsiveness (Lemanske and Busse, 2010). The key features of asthma include chronic inflammation of the airway, high responsiveness of the airway to a variety of stimulators, limited variable reversible flow, and a series of changes with the course of the disease, namely, airway reconstruction (Çalışkan et al., 2013). Asthma is one of the most common chronic diseases in the world, with about 300 million people worldwide and about 45 million asthma patients in China, and there is a trend year by year. Epidemiological studies have shown that early exposure to microbes may determine the composition of the microbiome, which can help prevent allergies or cause the development of asthma. Asthma had been demonstrated to be closely associated with microbes by a number of research (Gilstrap and Kraft, 2013). In this section, though the there is implementation of our model to infer the novel asthma-related microbes, we published evidence for the top 10 potential asthma-related microbes predicted by MSBMFHMDA in **Table 4**.

Inflammatory bowel disease (IBD) is a group of chronic non-specific intestinal inflammatory diseases that have no etiology, including ulcerative colitis and Crohn's disease (D'Aoust et al., 2017). In this paper, we selected IBD as one of our case studies to evaluate the performance of our model. As illustrated in the following **Table 5**, there are 10 out of these top 10 microbes predicted by MSBMFHMDA that have been substantiated to be associated with IBD.

Type 2 diabetes mellitus (T2D), also known as adult-onset diabetes, is characterized by a rise in blood sugar and a relative lack of insulin production because of a decline in the ability of insulin to help glucose enter cells for metabolism, a metabolic disorder resulting from a disorder of glucose metabolism (Furet et al., 2010). We took T2D as a case study for potential T2DM-related microbe prediction, and as illustrated in the following **Table 6**, 8 out of the top 10 predicted microbes were confirmed by experimental reports.

DISCUSSION AND CONCLUSION

Since the application of traditional experimental methods to identify disease-associated microbes is time consuming and expensive, the calculation approach of MSBMFHMDA was put forward. Our model provides an effective scheme for dynamically integrating multiple similarities and extracting useful features to infer potential microbe–disease associations. The non-negative constraint in the model also ensures that the predicted scores of associations are non-negative. The computational results demonstrate that MSBMFHMDA has good performances for microbe–disease association prediction.

However, our model has two limitations. First, there are only 450 known microbe–disease associations, which accounts for a very small proportion of human microbe diseases. This may result in less comprehensive for prediction. Second, our method involves non-convex optimization, which leads to the local optimal solutions instead of the global optimal solution. In the future, we will reform predictive tasks based on the HMDAD record additional entries whether the quantity of microbial population is increased or decreased in the reported cases.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/**Supplementary Material**.

AUTHOR CONTRIBUTIONS

XY and LW conceived and designed the study. XY, ZC, and LK obtained and processed the datasets. XY and LK wrote the paper. LW and LK provided suggestions and supervised the research.

FUNDING

The research is funded by the National Natural Science Foundation of China (No. 61873221) the Hunan Province Science and Technology Project Funds (2018TP1036) and the Natural Science Foundation of Hunan Province (No. 2019JJ70010).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.754425/full#supplementary-material

REFERENCES

Çalışkan, M., Bochkov, Y. A., Kreiner-Møller, E., Bønnelykke, K., Stein, M. M., Du, G., et al. (2013). Rhinovirus Wheezing Illness and Genetic Risk of Childhood-Onset Asthma. *N. Engl. J. Med.* 368, 1398–1407. doi:10.1056/NEJMoa1211592

Chen, X., Huang, Y.-A., You, Z.-H., Yan, G.-Y., and Wang, X.-S. (2016a). A Novel Approach Based on KATZ Measure to Predict Associations of Human Microbiota with Non-infectious Diseases. *Bioinformatics* 33, 733–739. doi:10.1093/bioinformatics/btw715

Chen, X., Liu, M.-X., and Yan, G.-Y. (2012). Drug-target Interaction Prediction by Random Walk on the Heterogeneous Network. *Mol. Biosyst.* 8, 1970. doi:10.1039/c2mb00002d

Chen, X., Sun, Y.-Z., Guan, N.-N., Qu, J., Huang, Z.-A., Zhu, Z.-X., et al. (2019). Computational Models for lncRNA Function Prediction and Functional Similarity Calculation. *Brief. Funct. Genomics* 18, 58–82. doi:10.1093/bfgp/ely031

Chen, X., Wang, L., Qu, J., Guan, N.-N., and Li, J.-Q. (2018a). Predicting miRNA-Disease Association Based on Inductive Matrix Completion. *Bioinformatics* 34, 4256–4265. doi:10.1093/bioinformatics/bty503

Chen, X., Xie, D., Wang, L., Zhao, Q., You, Z.-H., and Liu, H. (2018b). BNPMDA: Bipartite Network Projection for MiRNA-Disease Association Prediction. *Bioinformatics* 34, 3178–3186. doi:10.1093/bioinformatics/bty333

Chen, X., Yan, C. C., Zhang, X., and You, Z.-H. (2016b). Long Non-coding RNAs and Complex Diseases: from Experimental Results to Computational Models. *Brief. Bioinform.* 18, 558–576. doi:10.1093/bib/bbw060

Chen, X., and Yan, G.-Y. (2013). Novel Human lncRNA-Disease Association Inference Based on lncRNA Expression Profiles. *Bioinformatics* 29, 2617–2624. doi:10.1093/bioinformatics/btt426

Chen, X., and Yan, G.-Y. (2015). Semi-supervised Learning for Potential Human microRNA-Disease Associations Inference. *Sci. Rep.* 4, 5501. doi:10.1038/srep05501

Cui, Z., Gao, Y.-L., Liu, J.-X., Wang, J., Shang, J., and Dai, L.-Y. (2019). The Computational Prediction of Drug-Disease Interactions Using the Dual-Network L2,1-CMF Method. *BMC Bioinformatics* 20, 5. doi:10.1186/s12859-018-2575-6

D'Aoust, J., Battat, R., and Bessissow, T. (2017). Management of Inflammatory Bowel Disease withClostridium Difficileinfection. *Wjg* 23, 4986. doi:10.3748/wjg.v23.i27.4986

Furet, J.-P., Kong, L.-C., Tap, J., Poitou, C., Basdevant, A., Bouillot, J.-L., et al. (2010). Differential Adaptation of Human Gut Microbiota to Bariatric Surgery-Induced Weight Loss: Links with Metabolic and Low-Grade Inflammation Markers. *Diabetes* 59, 3049–3057. doi:10.2337/db10-0253

Gilstrap, D. L., and Kraft, M. (2013). Asthma and the Host-Microbe Interaction. *J. Allergy Clin. Immunol.* 131, 1449–1450. doi:10.1016/j.jaci.2013.03.004

Hawn, T. R., Day, T. A., Scriba, T. J., Hatherill, M., Hanekom, W. A., Evans, T. G., et al. (2014). Tuberculosis Vaccines and Prevention of Infection. *Microbiol. Mol. Biol. Rev.* 78, 650–671. doi:10.1128/MMBR.00021-14

Hendricks, K. A., Wright, M. E., Shadomy, S. V., Bradley, J. S., Morrow, M. G., Pavia, A. T., et al. (2014). Centers for Disease Control and Prevention Expert Panel Meetings on Prevention and Treatment of Anthrax in Adults. *Emerg. Infect. Dis.* 20. doi:10.3201/eid2002.130687

Huang, D.-S., and Du, J.-X. (2008). A Constructive Hybrid Structure Optimization Methodology for Radial Basis Probabilistic Neural Networks. *IEEE Trans. Neural Netw.* 19, 2099–2115. doi:10.1109/TNN.2008.2004370

Huang, D.-S. (1999). RADIAL BASIS PROBABILISTIC NEURAL NETWORKS: MODEL AND APPLICATION. *Int. J. Patt. Recogn. Artif. Intell.* 13, 1083–1101. doi:10.1142/S0218001499000604

Huang, D. (1996). Generalization Capabilities of Feedforward Neural Networks for Pattern Recognition. *J. Beijing Inst. Technol. Engl. Ed.* 5.

Lemanske, R. F., and Busse, W. W. (2010). Asthma: Clinical Expression and Molecular Mechanisms. *J. Allergy Clin. Immunol.* 125, S95–S102. doi:10.1016/j.jaci.2009.10.047

Liu, Y., Wang, S.-L., and Zhang, J.-F. (2018). Prediction of Microbe-Disease Associations by Graph Regularized Non-negative Matrix Factorization. *J. Comput. Biol.* 25, 1385–1394. doi:10.1089/cmb.2018.0072

Ma, W., Zhang, L., Zeng, P., Huang, C., Li, J., Geng, B., et al. (2017). An Analysis of Human Microbe-Disease Associations. *Brief. Bioinform.* 18, 85–97. doi:10.1093/bib/bbw005

Peng, L.-H., Yin, J., Zhou, L., Liu, M.-X., and Zhao, Y. (2018). Human Microbe-Disease Association Prediction Based on Adaptive Boosting. *Front. Microbiol.* 9, 2440. doi:10.3389/fmicb.2018.02440

Qiu, Y., Ching, W.-K., and Zou, Q. (2021a). Matrix Factorization-Based Data Fusion for the Prediction of RNA-Binding Proteins and Alternative Splicing Event Associations during Epithelial-Mesenchymal Transition. *Brief. Bioinform.*, bbab332. doi:10.1093/bib/bbab332

Qiu, Y., Ching, W.-K., and Zou, Q. (2021b). Prediction of RNA-Binding Protein and Alternative Splicing Event Associations during Epithelial-Mesenchymal Transition Based on Inductive Matrix Completion. *Brief. Bioinform.* 22, bbaa440. doi:10.1093/bib/bbaa440

Qu, J., Zhao, Y., and Yin, J. (2019). Identification and Analysis of Human Microbe-Disease Associations by Matrix Decomposition and Label Propagation. *Front. Microbiol.* 10, 291. doi:10.3389/fmicb.2019.00291

Salton, G., Wong, A., and Yang, C. S. (1975). A Vector Space Model for Automatic Indexing. *Commun. ACM* 18, 613–620. doi:10.1145/361219.361220

Shang, F., Cheng, J., Liu, Y., Luo, Z.-Q., and Lin, Z. (2018). Bilinear Factor Matrix Norm Minimization for Robust PCA: Algorithms and Applications. ArXiv181005186. *Cs Math. Stat.* Available at: http://arxiv.org/abs/1810.05186 (Accessed August 5, 2021).

Shen, X., Zhu, H., Jiang, X., Hu, X., and Yang, J. (2018). "A Novel Approach Based on Bi-random Walk to Predict Microbe-Disease Associations," in *Intelligent Computing Methodologies*. Editors D.-S. Huang, M. M. Gromiha, K. Han, and A. Hussain (Cham: Springer International Publishing)), 746–752. doi:10.1007/978-3-319-95957-3_78

Shen, Z., Jiang, Z., and Bao, W. (2017). "CMFHMDA: Collaborative Matrix Factorization for Human Microbe-Disease Association Prediction," in *Intelligent Computing Theories and Application*. Editors D.-S. Huang, K.-H. Jo, and J. C. Figueroa-García (Cham: Springer International Publishing)), 261–269. doi:10.1007/978-3-319-63312-1_24

Shi, J.-Y., Huang, H., Zhang, Y.-N., Cao, J.-B., and Yiu, S.-M. (2018). BMCMDA: a Novel Model for Predicting Human Microbe-Disease Associations via Binary Matrix Completion. *BMC Bioinformatics* 19, 281. doi:10.1186/s12859-018-2274-3

The Human Microbiome Project Consortium (2012). A Framework for Human Microbiome Research. *Nature* 486, 215–221. doi:10.1038/nature11209

van Laarhoven, T., Nabuurs, S. B., and Marchiori, E. (2011). Gaussian Interaction Profile Kernels for Predicting Drug-Target Interaction. *Bioinformatics* 27, 3036–3043. doi:10.1093/bioinformatics/btr500

Wang, J. J.-Y., and Gao, X. (2015). Max-min Distance Nonnegative Matrix Factorization. *Neural Networks* 61, 75–84. doi:10.1016/j.neunet.2014.10.006

Wang, L., Wang, Y., Li, H., Feng, X., Yuan, D., and Yang, J. (2019). A Bidirectional Label Propagation Based Computational Model for Potential Microbe-Disease Association Prediction. *Front. Microbiol.* 10, 684. doi:10.3389/fmicb.2019.00684

Wen, Z., Yan, C., Duan, G., Li, S., Wu, F.-X., and Wang, J. (2021). A Survey on Predicting Microbe-Disease Associations: Biological Data and Computational Methods. *Brief. Bioinform.* 22. doi:10.1093/bib/bbaa157

Wheeler, D. L., Barrett, T., Benson, D. A., Bryant, S. H., Canese, K., Chetvernin, V., et al. (2007). Database Resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* 36, D13–D21. doi:10.1093/nar/gkm1000

Xie, G., Meng, T., Luo, Y., and Liu, Z. (2019). SKF-LDA: Similarity Kernel Fusion for Predicting lncRNA-Disease Association. *Mol. Ther. - Nucleic Acids* 18, 45–55. doi:10.1016/j.omtn.2019.07.022

Xuan, Z., Li, J., Yu, J., Feng, X., Zhao, B., and Wang, L. (2019). A Probabilistic Matrix Factorization Method for Identifying lncRNA-Disease Associations. *Genes* 10, 126. doi:10.3390/genes10020126

Yan, C., Duan, G., Wu, F.-X., Pan, Y., and Wang, J. (2021). MCHMDA:Predicting Microbe-Disease Associations Based on Similarities and Low-Rank Matrix Completion. *Ieee/acm Trans. Comput. Biol. Bioinf.* 18, 611–620. doi:10.1109/tcbb.2019.2926716

Yang, M., Wu, G., Zhao, Q., Li, Y., and Wang, J. (2021). Computational Drug Repositioning Based on Multi-Similarities Bilinear Matrix Factorization. *Brief. Bioinform.* 22. doi:10.1093/bib/bbaa267

You, Z.-H., Huang, Z.-A., Zhu, Z., Yan, G.-Y., Li, Z.-W., Wen, Z., et al. (2017). PBMDA: A Novel and Effective Path-Based Computational Model for miRNA-Disease Association Prediction. *PLOS Comput. Biol.* 13, e1005455. doi:10.1371/journal.pcbi.1005455

Yu, J., Ping, P., Wang, L., Kuang, L., Li, X., and Wu, Z. (2018). A Novel Probability Model for LncRNA-Disease Association Prediction Based on the Naïve Bayesian Classifier. *Genes* 9, 345. doi:10.3390/genes9070345

Zheng, X., Ding, H., Mamitsuka, H., and Zhu, S. (2013). "Collaborative Matrix Factorization with Multiple Similarities for Predicting Drug-Target Interactions," in Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining (Chicago Illinois USA: ACM)), 1025–1033. doi:10.1145/2487575.2487670

Zhou, X., Menche, J., Barabási, A.-L., and Sharma, A. (2014). Human Symptoms-Disease Network. *Nat. Commun.* 5, 4212. doi:10.1038/ncomms5212

Zou, S., Zhang, J., and Zhang, Z. (2017). A Novel Approach for Predicting Microbe-Disease Associations by Bi-random Walk on the Heterogeneous Network. *PLOS ONE* 12, e0184394. doi:10.1371/journal.pone.0184394

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

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.