OXFORD

Briefings in Bioinformatics, 2025, **26(2)**, bbaf075

https://doi.org/10.1093/bib/bbaf075

Problem Solving Protocol

Ensemble learning based on matrix completion improves microbe-disease association prediction

Hailin Chen* and Kuan Chen

School of Information and Software Engineering, East China Jiaotong University, No. 808, Shuanggangdong Street, Nanchang 330013, China

*Corresponding author. School of Information and Software Engineering, East China Jiaotong University, Nanchang 330013, China. E-mail: chenhailin@ecjtu.edu.cn

Abstract

Microbes have a profound impact on human health. Identifying disease-associated microbes would provide helpful guidance for drug development and disease treatment. Through an enormous experimental effort, limited disease-associated microbes have been determined. Accurate computational approaches are needed to predict potential microbe-disease associations for biomedical screening. In this study, we present an ensemble learning framework entitled SABMDA to improve microbe-disease association inference. We first integrate multi-source of information from both microbes and diseases, and develop two matrix completion algorithms to predict microbe-disease associations successively. Ablation tests show combining the two matrix completion algorithms can receive better prediction performance. Moreover, comprehensive experiments, including cross-validations and independent test, demonstrate that SABMDA outperforms seven recent baseline methods significantly. Finally, we apply SABMDA to three diseases to predict their associated microbes, and results show SABMDA's remarkable prediction ability in real situations.

Keywords: microbe-disease association; ensemble learning; matrix completion

Introduction

Microbes widely exist on our planet, including in oceans, soils, and the human body [1]. It is estimated that the human body hosts approximately 350 trillion microbial cells [2]. With recent advances in sequencing and new bioinformatics development, significant progress has been made in revealing how microbiome composition and function affect human health. For example, studies showed that changes in the composition of human microbiota might impact immunological and pathological conditions [3–5]. Additionally, it has been discovered that hepatic metabolism was regulated by microbiota in the liver through decreasing energy expenditure and promoting adiposity [6]. Because of their fundamental roles in human health, accumulating studies have been conducted to elucidate the relationships between microbes and various types of diseases (see review [7] for more details).

Biomedical efforts to discover disease-associated microbes are often time-consuming and costly. Computational approaches to inferring potential microbe-disease associations would therefore bring benefits to the scientific communities. So far, developing algorithms for microbe-disease association prediction has aroused enormous interest in bioinformatics field [8]. Generally, these computational methods apply Random Walk [9–11], bipartite local models (BLMs) [12–14], matrix factorization [15–17], and machine learning [18, 19] for association prediction. These methods prioritize the potential associations according to their received scores based on graph theory, or consider the association prediction as a classification or regression problem in machine learning.

Even though, increasing prediction accuracy has gradually been received from the above computational methods. Some shortcomings should be mentioned. For example, the graph-based algorithms are sensitive to noise and spareness in data, which can lead to misleading predictions or severely impact performance. For the machine learning methods, selecting relevant features from high-dimensional biomedical data can be difficult. Poor feature selection can lead to suboptimal prediction performance.

More recently, with the rapid development of molecular biology science, new biomedical data about microbes and diseases is continuously emerging. The heterogeneous data provides complementary information while may contain noise. Integrating the biomedical information from different sources would improve the accuracy of microbe-disease association prediction. Meanwhile, improved prediction performance is required to offer useful guidance for biomedical researchers. More reliable prediction algorithms therefore need to be developed with the fast advance of modern computer science.

In this study, we present an ensemble learning method entitled SABMDA based on matrix completion to improve microbe-disease association prediction. Specifically, SABMDA first fuses multiple biomedical information of microbes and diseases to form a microbe-disease matrix. It then applies a singular value thresholding (SVT) algorithm to complete the original microbe-disease matrix. We finally use a bounded nuclear norm regularization (BNNR) algorithm with constraints to predict microbe-disease associations. By 5-CV, 10-CV, and independent validation tests, SABMDA receives the best prediction performance when compared to seven baseline methods. In addition, case studies

Received: November 8, 2024. **Revised:** January 18, 2025. **Accepted:** February 10, 2025

are conducted and results show that SABMDA exhibits reliable inference ability in real situations. In summary, the excellent performance of SABMDA suggests that it is a powerful and effective computational tool for inferring new microbe-disease associations.

Materials and methods

Data preparation

We first download the benchmark dataset from reference [19], in which 4499 experimentally confirmed microbe-disease associations were collected and four categories of similarities of both microbe-microbe and disease-disease were calculated. For the 4499 associations, there exist 1177 microbes and 134 diseases. For the four kinds of similarities in microbes, we refer to as functional similarity (FS), cosine similarity (COS_MS), Gaussian interaction profile similarity (GIP_MS), and sigmoid kernel function similarity (SIG_MS). For disease similarities, semantic similarity (DS), cosine similarity (COS_DS), Gaussian interaction profile similarity (GIP_DS), and sigmoid kernel function similarity (SIG_DS) were computed.

We then fuse the four similarity matrices of microbes and diseases as follows:

$$SM = \frac{FS + COS\_MS + GIP\_MS + SIG\_MS}{4} \tag{1}$$

$$SD = \frac{DS + COS\_DS + GIP\_DS + SIG\_DS}{4} \tag{2}$$

where SM denotes the fused microbe similarity matrix and SD represents the fused disease similarity matrix. Finally, we integrate the two fused similarity matrices with the microbe-disease association matrix A' to obtain a new matrix X:

$$X = \begin{bmatrix} SD & A'^{T} \\ A' & SM \end{bmatrix} \tag{3}$$

The framework of SABMDA

The workflow of SABMDA for microbe-disease association predictions is shown in Fig. 1. After similarity fusion, SABMDA combines the similarity matrices with the adjacency matrices, and performs matrix completion using a SVT algorithm. Then, a BNNR algorithm is applied to further optimize the complemented matrices with meta-type integration, and finally a score matrix for predicting microbe-disease associations is obtained. We prioritize the scores to screen potential associations.

Specifically, we consider the microbe-disease association prediction as a matrix completion problem and first apply a SVT algorithm to solve this problem. The algorithm was previously proposed as a solution to the famous Netflix problem [20]. It hierarchizes the data with different features and finds the optimal threshold value based on the specificity of each feature to achieve an accurate classification of the relevant data.

We update the matrix values by iterations, and in each iteration a matrix $X_i \in \mathbb{R}^{N_{m+d} \times N_{m+d}}$ (i denotes the number of current iterations) is generated. When the iteration ends, a matrix X_n (n denotes the number of final iterations) is obtained showcasing all microbe-disease association scores. In order to ensure that the association score matrix X_A in X_n is close to the score in the adjacency matrix A, the following optimization problem needs to be solved:

$$\begin{aligned} &\min_{X_A} f_t(X_A) \\ &s.t. P_\Omega(X_A) = P_\Omega(A) \end{aligned} \tag{4}$$

where X_A is the training matrix after removing the validation component, and P_Ω is the orthogonal projector over the span of the matrix that vanishes outside Ω. The $(i, j)_{th}$ component of $P_\Omega(X_A)$ is equal to X(i, j), if $(i,j) \in \Omega$ and zero otherwise. $f_t(X_A)$ is a nonlinear function of X_A and is defined in the following form:

$$f_t(X_A) = \tau \|X_A\|_* + \frac{1}{2}\|X_A\|_F^2 \tag{5}$$

where $\|X_A\|_*$ is the sum of singular values of XA, $\|X_A\|_F$ denotes the Frobenius form of X_A, which can also be denoted as $\|X_A\|_F = \sqrt{\sum_{i=1}^{n_d}\sum_{j=1}^{n_m} X_A(i,j)^2}$, and τ is a threshold value.

Then, we introduce the SVT operator and consider the singular value decomposition of a matrix $X \in \mathbb{R}^{N_{m+d} \times N_{m+d}}$ of rank r. The definition is as follows:

$$X = U\Sigma V^*, \Sigma = \mathrm{diag}\left(\{\sigma_i\}_{1 \le i \le r}\right) \tag{6}$$

where U and V represent $N_{m+d} \times r$ matrices with orthogonal columns and σ_i is singular and greater than zero. For each $\tau \ge 0$, we introduce the soft threshold operator D_τ defined as follows:

$$D_\tau(X) := UD_\tau(\Sigma)V^*, D_\tau(\Sigma) = \mathrm{diag}\left(\{\sigma_i - \tau\}_+\right) \tag{7}$$

where $\{\sigma_i - \tau\}_+$ represents the positive part of $\{\sigma_i - \tau\}$. This operation is able to apply the soft threshold rule to the singular values of X, effectively shrinking these singular values to zero.

According to reference [21], SVT can be optimized using the Lagrange multiplier method, and the Lagrange multiplier Y can be obtained as follows:

$$L(X, Y) = f_t(X) + < Y, P_\Omega(M) - P_\Omega(X) > \tag{8}$$

where M is defined as $M = \begin{bmatrix} SD & A^T \\ A & SM \end{bmatrix}$. In each iteration, we apply two key steps from Uzawa's algorithm [22]. The first one is to update X with Y:

$$X^k = D_\tau\left(Y^{k-1}\right) \tag{9}$$

Then we use X to update Y:

$$Y^k = Y^{k-1} + \delta_k\left(M - X^k\right) \tag{10}$$

where Y^0 is the zero matrix [23], and δ_k is the step size. We assume that the iterations converge to a unique solution when $0 < \delta_k < 2$. We show the best performance of our model can be received when $\delta_k = 0.1$ in parametric experiments. Subsequently, we set the iteration period n to limit the maximum number of iterations to avoid infinite loops, and details about the setting of n will be described in the Results part. Finally, we obtain the matrix X_n, which is then sigmoid, normalized to obtain X_n' using the following equation:

$$X_n' = \frac{1}{1 + e^{-X_n}} \tag{11}$$

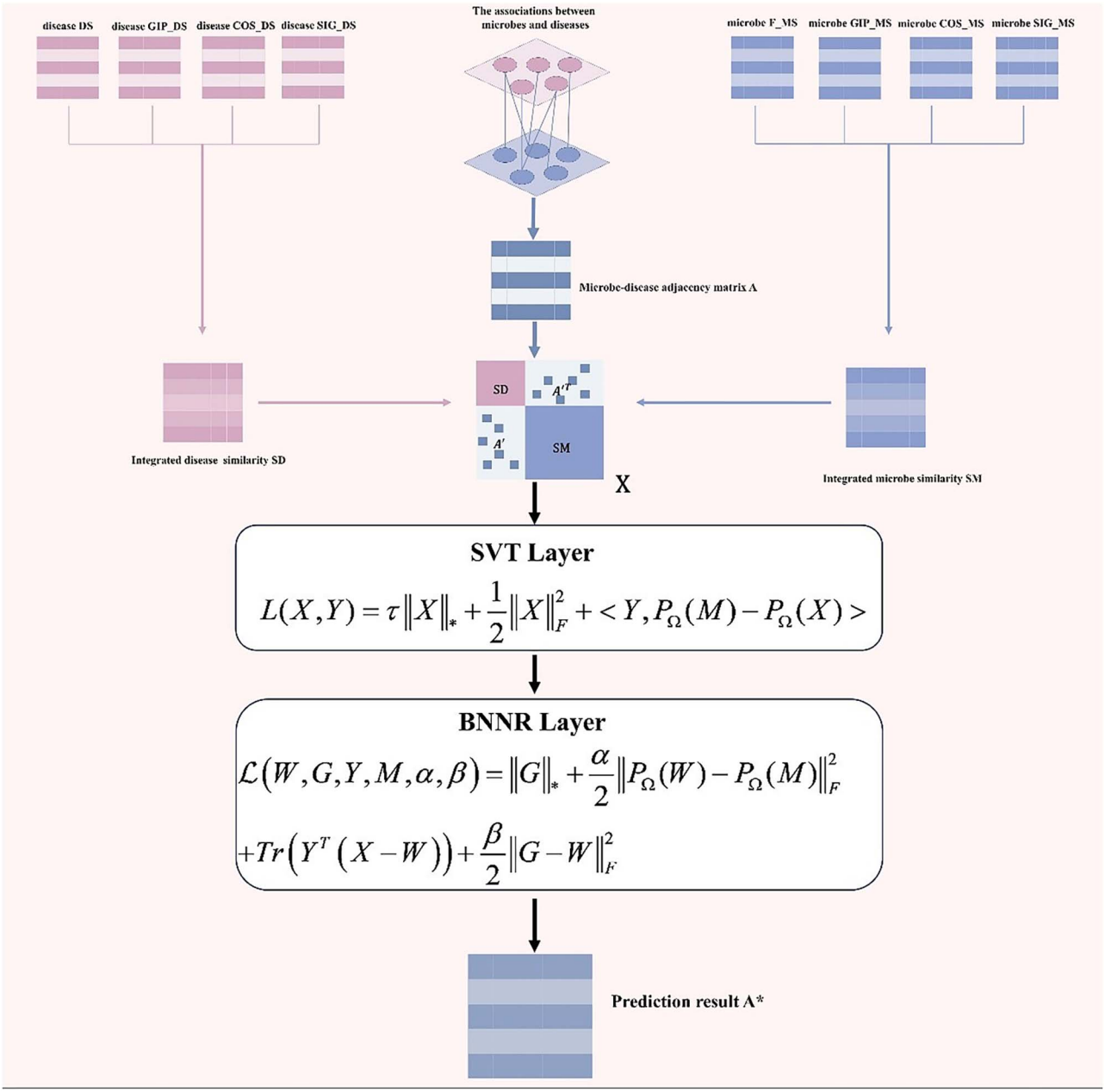

Figure 1. The workflow of SABMDA for microbe-disease association prediction.

We extract the score matrix X_A from the corresponding position in X_n'. We further construct a heterogeneous network G that integrates disease-disease similarity network, microbe-disease network, and microbe-microbe similarity network, as follows:

$$G=\begin{bmatrix} SD & X_A^T \\ X_A & SM \end{bmatrix} \tag{12}$$

As the elements in the microbe similarity matrix SM and the disease similarity matrix SD are in the interval [0,1], and the elements in X_A are also in the interval [0,1], we expect the predicted values of the unknown associations to be in the interval [0,1] too. Therefore, we add boundary treaties to our model to ensure that the elements to be predicted are also in the interval [0,1].

In addition, since there is a large amount of noise in the data, especially when measuring similarities, our proposed model should effectively tolerate this noise. We therefore redefine the model as:

$$\begin{aligned} &\min_G \|G\|_* \\ &s.t. \|P_\Omega(G)-P_\Omega(M)\|_F \le \varepsilon \end{aligned} \tag{13}$$

Where ε measures the level of noise. Meanwhile, because the level of noise is unknown, it is not easy to calculate the effective noise level. Therefore, in order to solve the above problems, inspired by references [24, 25], we further optimize the score matrix X_A with meta-level type integration. The model is defined as follows:

$$\begin{aligned} &\min_G \|G\|_* + \tfrac{\alpha}{2}\|P_\Omega(G)-P_\Omega(M)\|_F^2 \\ &s.t. 0 \le G \le 1 \end{aligned} \tag{14}$$

where α is the parameter that balances the kernel paradigm and the error term. To make sure all elements in G belong to the interval [0,1], we use ADMM [26] to solve this problem.

Before using the ADMM framework for problem solving, we introduce an auxiliary function W for optimization:

$$\begin{aligned} &\min_{G} \|G\|_* + \tfrac{\alpha}{2}\|P_\Omega(W) - P_\Omega(M)\|_F^2 \\ &s.t. G = W, 0 \le W \le 1 \end{aligned} \tag{15}$$

Thus, the augmented Lagrangian function can be written in the following form:

$$\begin{aligned} &\mathcal{L}(W, G, Y, M, \alpha, \beta) = \|G\|_* + \tfrac{\alpha}{2}\|P_\Omega(W) - P_\Omega(M)\|_F^2 \\ &+ Tr\left(Y^T(X - W)\right) + \tfrac{\beta}{2}\|G - W\|_F^2 \end{aligned} \tag{16}$$

where Y is the Lagrange multiplier and β is the penalty parameter and is greater than 0. After the Kth time, the model computes W_{k+1}, G_{k+1}, and Y_{k+1}.

ADMM is then applied to iteratively update W_{k+1}, G_{k+1}, and Y_{k+1}. Their iterative procedures are denoted as follows:

$$W_{k+1} = \arg\min_{0 \le W \le 1} \mathcal{L}(W, G, Y, M, \alpha, \beta) \tag{17}$$

$$G_{k+1} = \arg\min_{G} \mathcal{L}(W, G, Y, M, \alpha, \beta) \tag{18}$$

$$Y_{k+1} = Y_k + \beta(G_{k+1} - W_{k+1}) \tag{19}$$

Based on Equation (17), we can obtain the optimal solution W* for W_{k+1}, which is computed as follows:

$$\begin{aligned} W^* = &\left(\tfrac{1}{\beta}Y_k + \tfrac{\alpha}{\beta}P_\Omega(M) + G_k\right) - \tfrac{\alpha}{\alpha+\beta} \\ &\left(\tfrac{1}{\beta}Y_k + \tfrac{\alpha}{\beta}P_\Omega(M) + G_k\right) \end{aligned} \tag{20}$$

According to Equation (18), G_{k+1} can be calculated by the following equation:

$$G_{k+1} = D_{\frac{1}{\beta}}\left(W_{k+1} - \frac{1}{\beta}Y_k\right) \tag{21}$$

Finally, we can get the microbe-disease score matrix X_A' from G_{k+1} after iterations. We prioritize the received scores for association prediction.

Results

Experimental setting

To evaluate the prediction ability of SABMDA, we perform 5-fold cross-validation (5-CV), 10-fold cross-validation (10-CV), and independent test based on the benchmark dataset. For 5-CV, we randomly divide all the microbe-disease associations into 5 equal portions, of which 4 portions are used to train the model, and the remaining 1 portion is used for testing. We take the similar steps in 10-CV. For the independent test, where the microbe-disease association matrix is divided into training, testing, and validation sets by rows (diseases) according to the ratio of 8:1:1, in which 8 parts are for the training set, 1 part is for the testing set and 1 part is for the validation set. We calculate AUC (area under the ROC curve), AUPR (area under the Precision-Recall curve), Recall, Precision (Pre), Accuracy (ACC), and F1-score as the metrics for evaluating the performance of the model.

Parameter analysis

Since parameters exist in our method SABMDA, we set the initial values for the parameters based on references [27, 28], and subsequently empirically perform parameter sensitivity analysis for adjustment. We examine the effects of the threshold value τ, step size δ_k, iteration period n, as well as the parameter α that balances the kernel paradigm and the error term, and the penalty parameter β. While keeping all other parameters fixed, we conduct experiments on the benchmark dataset and evaluate the impact of the above parameters on the performance under 10-CV.

Firstly, we test the threshold value τ and step size δ_k. We take the threshold value τ as 10, 50, 100, and 150, and the step size δ_k as 0.1, 0.5, 1.0, and 1.5, respectively. The results are shown in Fig. 2 and Table 1. We discover that the AUC, AUPR, ACC, and F1-Score of our model will reach the best when the threshold value $\tau = 10$, and the step size $\delta_k = 0.1$.

Secondly, we carry out experiments on the iteration period n, and select its value from 50, 100, 200, 500, 1000, and 2000. Figure 3 and Table 2 show that the performance of our method will be optimal when the iteration period $n = 500$.

Finally, we analyze the effects of two parameters α and β. We take their values as 1.0, 10.0, 50.0, and 100.0, and the results are shown in Fig. 4 and Table 3, in which when $\alpha = 1.0$ and $\beta = 50.0$, the best performance can be received for our model.

To summarize, we finally set the parameters $\tau = 10$, $\delta_k = 0.1$, $n = 500$, $\alpha = 1.0$ and $\beta = 50.0$ in our model.

Ablation test

Our computational framework applies two matrix completion strategies (i.e. SVT and BNNR) for prediction. We therefore conduct ablation experiments based on 10-CV to investigate the impact of these components on model performance. Below are the three models we compare:

SABMDA-SVT model: we remove the SVT algorithm and predict microbe-disease associations only by the BNNR algorithm.

SABMDA-BNNR model: we remove the BNNR algorithm and make predictions only through the SVT algorithm.

SABMDA-SO model: we first perform the BNNR algorithm to generate the score matrix and then apply the SVT algorithm to further meta-type integration of the generated score matrix with the parameters kept unchanged.

It can be concluded from Fig. 5 that both the two matrix completion strategies are excellent in filling in the missing values in our study. After the first round of matrix completion, some of the missing values in the original matrix are recovered. Based on the new matrix, the second round of matrix completion makes more missing values recovered. Our ensemble learning framework SABMDA finally demonstrate superior performance.

Performance comparison with other methods

In this section, we compare our model SABMDA with the following baseline methods:

SGJMDA [29]: a method based on similarity fusion using graph convolution networks and jumping knowledge networks for microbe-disease association predictions.

DSAE_RF [19]: a method combining deep sparse autoencoder neural network (DSAE) and random forest (RF) for microbe-disease association predictions.

AMHMDA [30]: a method based on attention aware multi-view similarity networks and hypergraph learning for MiRNA-Disease Associations identification.

MHCLMDA [31]: a multihypergraph contrastive learning method for miRNA–disease association predictions.

MNNMDA [32]: a method to predict microbe-disease associations (MDAs) by applying a Matrix Nuclear Norm method into known microbe and disease data.

LRLSHMDA [13]: a method using Laplacian regularized least squares for human microbe–disease association predictions.

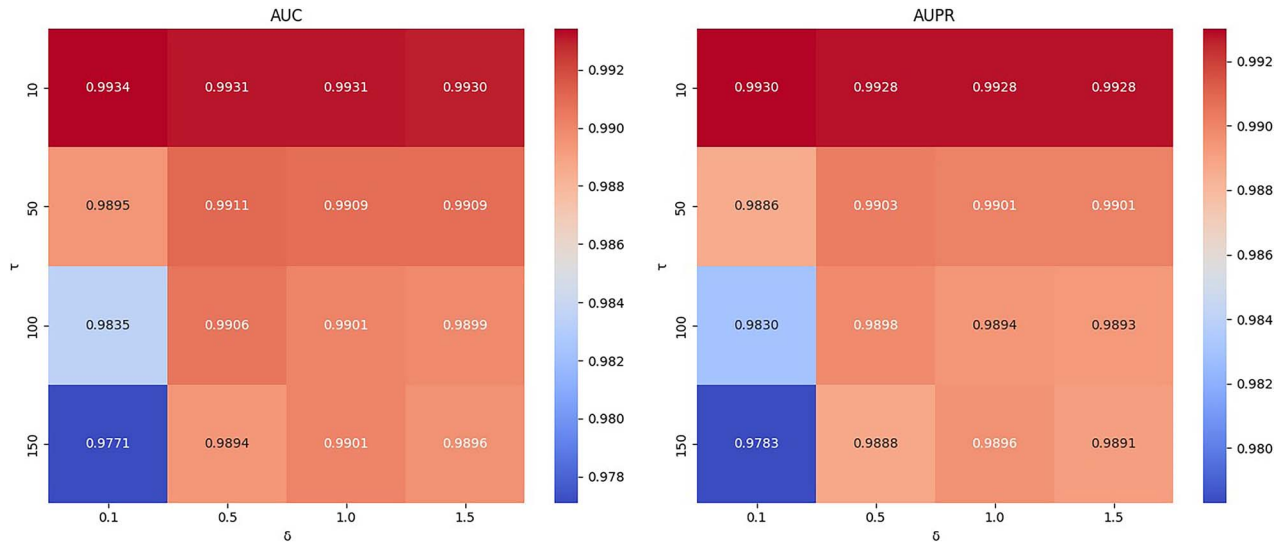

Figure 2. The effect of τ and δ_k on AUC and AUPR.

Table 1. The effect of τ and δ_k on model performance

	AUC	AUPR	ACC	Pre	Recall	F1-score
$\tau = 10, \delta_k = 0.1$	**0.9934**	**0.9930**	**0.9658**	0.9616	0.9706	**0.9660**
$\tau = 10, \delta_k = 0.5$	0.9931	0.9928	0.9655	0.9617	0.9697	0.9656
$\tau = 10, \delta_k = 1.0$	0.9931	0.9928	0.9655	**0.9618**	0.9698	0.9657
$\tau = 10, \delta_k = 1.5$	0.9930	0.9928	0.9653	0.9598	**0.9713**	0.9655
$\tau = 50, \delta_k = 0.1$	0.9895	0.9886	0.9563	0.9543	0.9586	0.9564
$\tau = 50, \delta_k = 0.5$	0.9911	0.9903	0.9596	0.9546	0.9653	0.9599
$\tau = 50, \delta_k = 1.0$	0.9909	0.9901	0.9588	0.9536	0.9648	0.9591
$\tau = 50, \delta_k = 1.5$	0.9909	0.9901	0.9587	0.9533	0.9649	0.9590
$\tau = 100, \delta_k = 0.1$	0.9835	0.9830	0.9436	0.9389	0.9495	0.9439
$\tau = 100, \delta_k = 0.5$	0.9906	0.9898	0.9579	0.9549	0.9615	0.9581
$\tau = 100, \delta_k = 1.0$	0.9901	0.9894	0.9571	0.9534	0.9613	0.9572
$\tau = 100, \delta_k = 1.5$	0.9899	0.9893	0.9569	0.9521	0.9624	0.9572
$\tau = 150, \delta_k = 0.1$	0.9771	0.9783	0.9328	0.9365	0.9290	0.9325
$\tau = 150, \delta_k = 0.5$	0.9894	0.9888	0.9538	0.9491	0.9593	0.9541
$\tau = 150, \delta_k = 1.0$	0.9901	0.9896	0.9562	0.9524	0.9604	0.9563
$\tau = 150, \delta_k = 1.5$	0.9896	0.9891	0.9556	0.9523	0.9595	0.9558

Note: The best results are marked in bold.

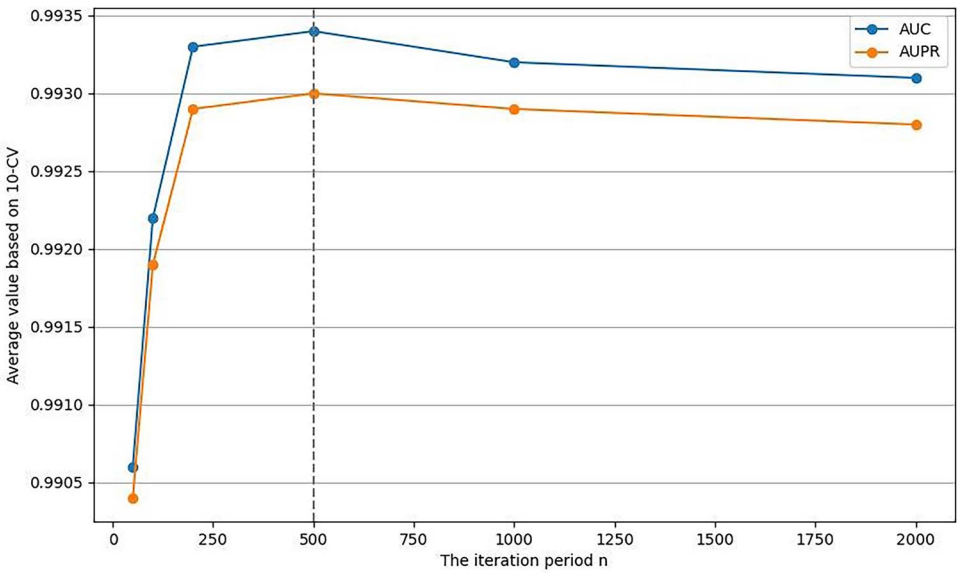

Figure 3. The effect of iteration period n in parameter analysis.

Table 2. The effect of iteration period n on model performance

	AUC	AUPR	ACC	Pre	Recall	F1-score
50	0.9906	0.9904	0.9589	0.9525	0.9662	0.9592
100	0.9922	0.9919	0.9631	0.9536	**0.9735**	0.9634
200	0.9933	0.9929	**0.9663**	0.9599	0.9733	**0.9665**
500	**0.9934**	**0.9930**	0.9658	**0.9616**	0.9706	0.9660
1000	0.9932	0.9928	0.9655	0.9594	0.9722	0.9657
2000	0.9931	0.9928	0.9654	0.9600	0.9713	0.9656

Note: The best results are marked in bold.

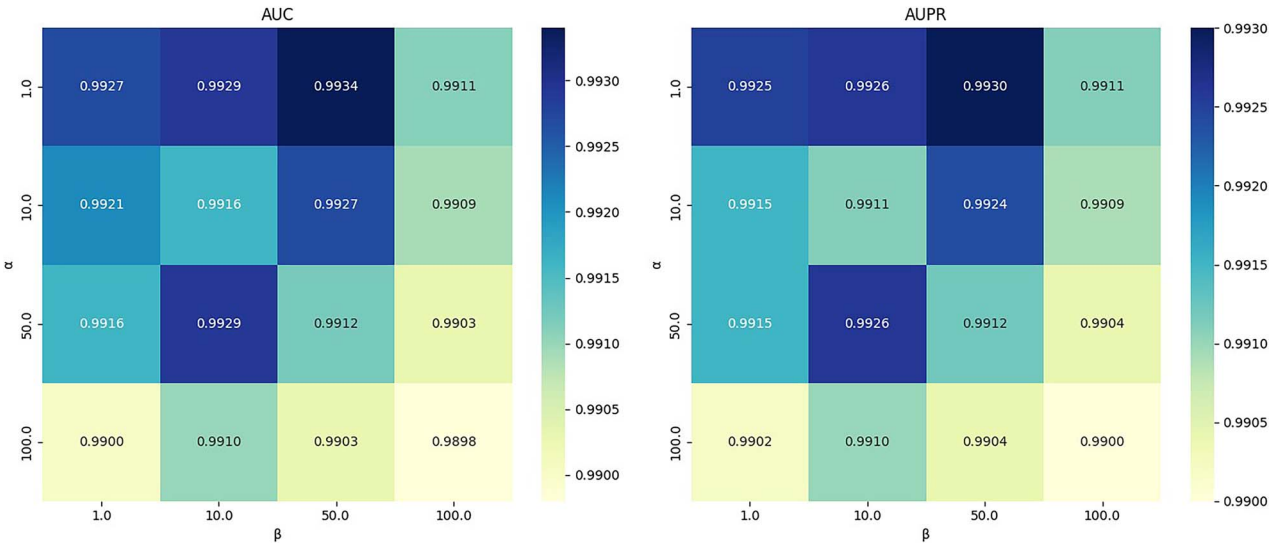

Figure 4. The effect of α and β on AUC and AUPR.

Table 3. The effect of α and β on model performance

	AUC	AUPR	ACC	Pre	Recall	F1-score
$\alpha = 1.0, \beta = 1.0$	0.9927	0.9925	0.9616	**0.9625**	0.9608	0.9616
$\alpha = 1.0, \beta = 10.0$	0.9929	0.9926	0.9631	0.9600	**0.9731**	0.9635
$\alpha = 1.0, \beta = 50.0$	**0.9934**	**0.9930**	**0.9658**	0.9616	0.9706	**0.9660**
$\alpha = 1.0, \beta = 100.0$	0.9911	0.9911	0.9614	0.9546	0.9688	0.9617
$\alpha = 10.0, \beta = 1.0$	0.9921	0.9919	0.9630	0.9536	0.9730	0.9623
$\alpha = 10.0, \beta = 10.0$	0.9916	0.9911	0.9604	0.9552	0.9667	0.9612
$\alpha = 10.0, \beta = 50.0$	0.9927	0.9924	0.9645	0.9585	0.9711	0.9647
$\alpha = 10.0, \beta = 100.0$	0.9909	0.9909	0.9608	0.9542	0.9682	0.9611
$\alpha = 50.0, \beta = 1.0$	0.9916	0.9915	0.9624	0.9534	0.9724	0.9628
$\alpha = 50.0, \beta = 10.0$	0.9929	0.9926	0.9647	0.9579	0.9722	0.9650
$\alpha = 50.0, \beta = 50.0$	0.9912	0.9912	0.9614	0.9541	0.9695	0.9617
$\alpha = 50.0, \beta = 100.0$	0.9903	0.9904	0.9590	0.9517	0.9673	0.9594
$\alpha = 100.0, \beta = 1.0$	0.9900	0.9902	0.9589	0.9531	0.9655	0.9592
$\alpha = 100.0, \beta = 10.0$	0.9910	0.9910	0.9611	0.9544	0.9684	0.9613
$\alpha = 100.0, \beta = 50.0$	0.9903	0.9904	0.9592	0.9515	0.9677	0.9595
$\alpha = 100.0, \beta = 100.0$	0.9898	0.9900	0.9583	0.9513	0.9662	0.9586

Note: The best results are marked in bold.

NTSHMDA [33]: a method to predict Human Microbe-Disease Associations based on Random Walk by Integrating Network Topological Similarity.

All methods are compared based on the same experimental setup. For the 5-CV experiment, we plot the ROC and PR curves in Fig. 6. The results show that SABMDA has the highest AUC and AUPR values, where the AUC value is 0.9919 and the AUPR value is 0.9920, which are 4.52% and 5.43% higher than the second-best method SGJMDA, respectively. The values of all the performance indicators for the 5-CV are shown in Table 4. We further calculate P-values (Table 5) based on the AUC and AUPR results received from the 5-CV experiments, which indicates the significant differences between our method and the other seven baseline methods. All the results show that SABMDA outperforms the seven state-of-the-art methods based on 5-CV.

For 10-CV, we plot ROC and PR curves in Fig. 7. The detailed results are listed in Table 6. As can be seen from Table 6, SABMDA outperforms the other seven methods in all assessment metrics, with an AUC value of 0.9934, which is 4.39% higher than SGJMDA, which has the second best results, and an AUPR value of 0.9930,

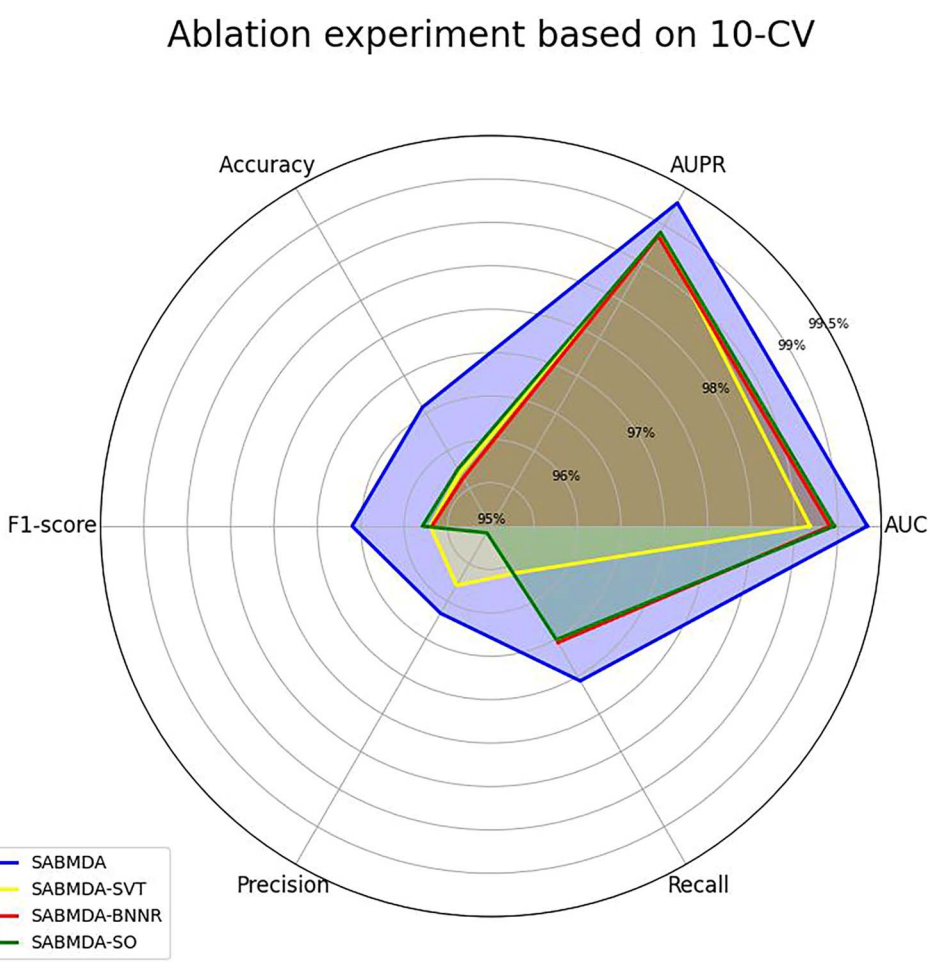

Figure 5. Results of ablation experiments based on 10-CV.

Table 4. Performance comparison based on 5-CV

Method	AUC	AUPR	ACC	Pre	Recall	F1-score
SABMDA	**0.9919**	**0.992**	**0.9614**	**0.9488**	**0.9648**	**0.9755**
SGJMDA	0.9467	0.9377	0.8851	0.8673	0.9099	0.8879
DSAE_RF	0.9238	0.9178	0.8484	0.849	0.8482	0.8481
AMHMDA	0.8883	0.8813	0.7902	0.8339	0.7313	0.775
MHCLMDA	0.8841	0.8763	0.7178	0.7787	0.8635	0.8187
MNNMDA	0.9107	0.9196	0.8886	0.902	0.8726	0.8867
LRLSHMDA	0.8233	0.7906	0.7617	0.7195	0.8606	0.7832
NTSHMDA	0.7972	0.772	0.7128	0.6646	0.8657	0.7507

Note: The best results are marked in bold.

Table 5. Statistical test results based on 5-CV between SABMDA and the other seven methods

	SGJMDA	DSAE_RF	AMHMDA	MHCLMDA	MNNMDA	LRLSHMDA	NTSHMDA
p-value based on AUC results	2.69×10^{-10}	1.00×10^{-7}	4.95×10^{-11}	6.62×10^{-9}	1.40×10^{-6}	1.78×10^{-10}	1.80×10^{-13}
p-value based on AUPR results	2.05×10^{-7}	9.11×10^{-9}	7.68×10^{-10}	1.90×10^{-7}	1.78×10^{-7}	1.83×10^{-9}	3.40×10^{-11}

which is 5.02% higher than SGJMDA, which has the second best results. Further statistical test results (Table 7) also indicate the significant differences between our method and the other seven baseline methods.

Independent test

To further validate the prediction performance of SABMDA, we conduct independent tests on our model SABMDA by dividing the microbe-disease association matrix by rows (diseases) into a training set, a test set, and a validation set, with a ratio of 8:1:1. We plot the ROC and PR curves of the independent tests in Fig. 8. In addition, the detailed results of all metrics are displayed in Table 8. It can be found that SABMDA receives scores of 0.8570, 0.8726, 0.8034, 0.7968, 0.8195, and 0.8065 for AUC, AUPR, ACC, Pre, Recall, and F1-Score, respectively. Compared with the other seven models, the other metrics are optimal, except for Recall, which does not reach the highest value. Taken together, these significant advantages highlight the effectiveness and excellence of SABMDA compared with existing methods.

Table 6. Performance comparison based on 10-CV

Method	AUC	AUPR	ACC	Pre	Recall	F1-score
SABMDA	**0.9934**	**0.993**	**0.9658**	**0.9616**	**0.9706**	**0.966**
SGJMDA	0.9495	0.9428	0.8901	0.8698	0.919	0.8933
DSAE_RF	0.9255	0.9199	0.848	0.8486	0.848	0.8478
AMHMDA	0.8922	0.8854	0.7955	0.8443	0.7274	0.7794
MHCLMDA	0.8817	0.8673	0.7295	0.7723	0.8844	0.8237
MNNMDA	0.9209	0.9347	0.892	0.8986	0.885	0.8914
LRLSHMDA	0.8292	0.7949	0.7747	0.7338	0.8628	0.7928
NTSHMDA	0.7962	0.7695	0.7148	0.6663	0.8762	0.7548

Note: The best results are marked in bold.

Table 7. Statistical test results based on 10-CV between SABMDA and the other seven methods

	SGJMDA	DSAE_RF	AMHMDA	MHCLMDA	MNNMDA	LRLSHMDA	NTSHMDA
p-value based on AUC results	2.11×10^{-13}	2.37×10^{-15}	8.71×10^{-25}	1.08×10^{-17}	1.53×10^{-13}	9.26×10^{-21}	5.77×10^{-20}
p-value based on AUPR results	2.81×10^{-13}	2.65×10^{-16}	3.67×10^{-22}	3.43×10^{-15}	9.53×10^{-16}	3.31×10^{-20}	1.30×10^{-19}

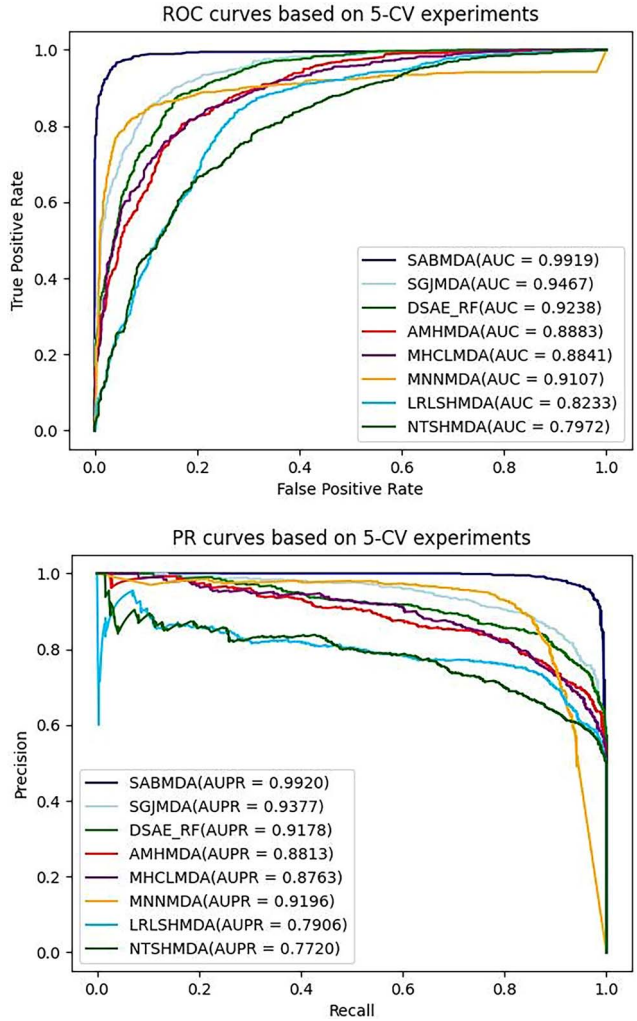

Figure 6. ROC and PR curves based on 5-CV experiments.

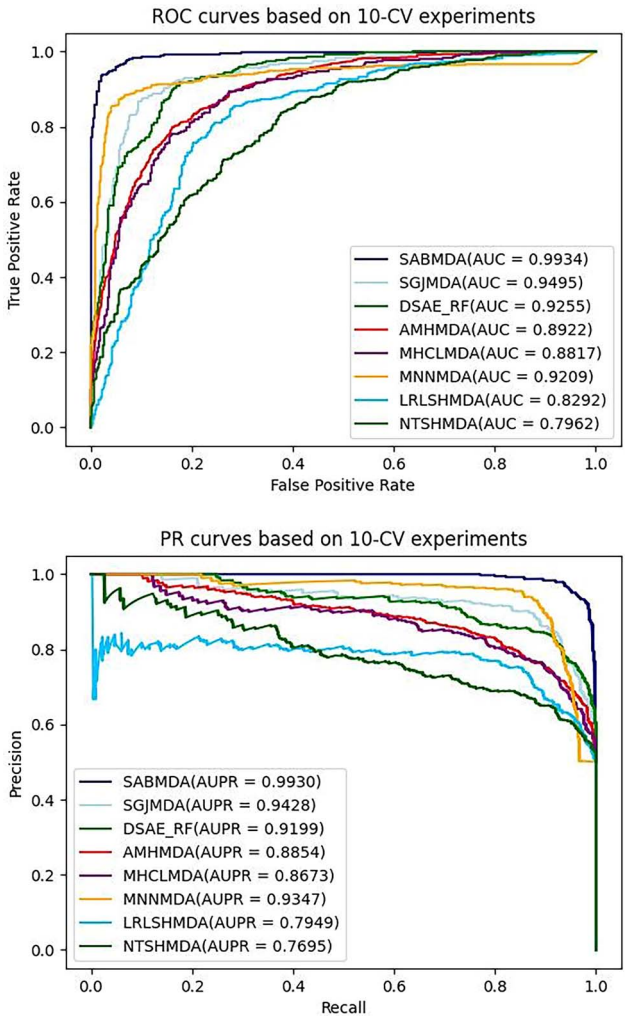

Figure 7. ROC and PR curves based on 10-CV experiments.

Table 8. Performance comparison based on independent test

Method	AUC	AUPR	ACC	Pre	Recall	F1-score
SABMDA	**0.857**	**0.8726**	**0.8034**	**0.7968**	**0.8195**	**0.8065**
SGJMDA	0.7842	0.7796	0.7043	0.6552	0.8799	0.7492
DSAE_RF	0.7796	0.8038	0.7271	0.7324	0.7271	0.7254
AMHMDA	0.7782	0.7847	0.6807	0.7919	0.6537	0.6005
MHCLMDA	0.6108	0.5912	0.5714	0.5565	**0.926**	0.693
MNNMDA	0.71	0.743	0.629	0.5907	0.8848	0.7053
LRLSHMDA	0.714	0.7754	0.6863	0.6997	0.7086	0.6948
NTSHMDA	0.6114	0.5711	0.5777	0.5449	0.9429	0.6906

Note: The best results are marked in bold.

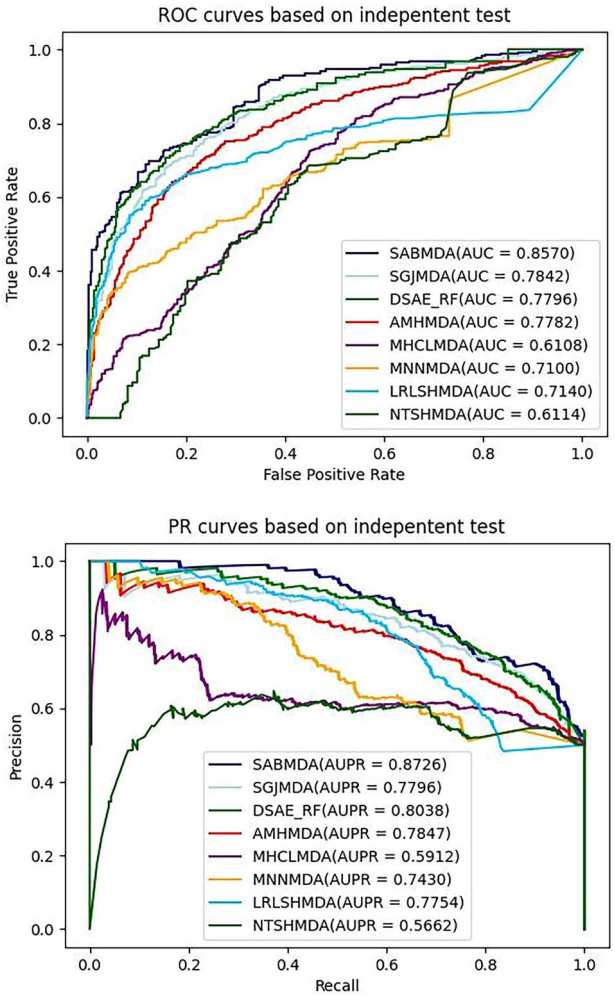

Figure 8. ROC and PR curves based on independent test.

Robustness analysis

To further evaluate the generalization ability of SABMDA for association prediction, we apply our model to the HMDD v3.2 dataset [34] for inference, and the results based on 10-CV are shown in Table 9. The results show that SABMDA also exhibits excellent prediction performance on other datasets, which validates SABMDA's wide application.

Case studies

In this section, we further test the prediction ability of SABMDA based on case studies. We first conduct the experiments by removing the association information of two specific diseases (i.e. *OBESITY* [35] and *ASTHMA* [36]) from the benchmark dataset and then apply our model to predict the microbes associated with the two diseases. We search the latest version of PubMed (https://pubmed.ncbi.nlm.nih.gov/) for confirmation, and the increased or decreased microbiota profile level is manually extracted from the related papers. The validation results are listed in Tables 10 and 11, respectively.

Moreover, we use the whole information in the benchmark dataset and then apply SABMDA to predict potential microbe-disease associations. We check the top 20 potential microbes for *CROHN'S DISEASE* [37] and the top 20 potential microbe-disease associations, and Tables 12 and 13 show the results of our experiments. The results of the four case studies indicate that SABMDA is an effective tool in predicting new microbe-disease associations.

Table 9. The performance of SABMDA on the HMDD v3.2 dataset

Dataset	AUC	AUPR	ACC	Pre	Recall	F1-score
HMDD v3.2	0.9475	0.9540	0.8885	0.8831	0.8961	0.8894

Table 10. The top 20 predicted microbes associated with *OBESITY*

Ranking	Microbe	Evidence	Description
1	*Streptococcus*	NA	NA
2	*Faecalibacterium*	NA	NA
3	*Haemophilus*	PMID:31976177	Increased
4	*Streptococcus mitis*	NA	NA
5	*Paraprevotella*	PMID:30525950	Increased
6	*Dialister*	PMID:32624568	Decreased
7	*Parabacteroides*	PMID:31530820	Increased
8	*Prevotella*	PMID:31024514	Decreased
9	*Akkermansia*	PMID:30810328	Decreased
10	*Roseburia*	PMID:34978141	Increased
11	*Streptococcus salivarius*	PMID:36264094	Decreased
12	*Streptococcus gordonii*	NA	NA
13	*Bifidobacterium*	PMID:29280312	Decreased
14	*Alloprevotella*	PMID:30611080	Decreased
15	*Ruminococcus*	PMID:31315227	Increased
16	*Megasphaera*	PMID:39033197	Increased
17	*Bacteroidetes*	PMID:17183312	NA
18	*Actinomyces*	PMID:35880087	Increased
19	*Fusobacterium*	PMID:29280312	Increased
20	*Veillonella*	PMID:31024514	Decreased

Note: NA indicates not available. Different conditions of obesity, such as childhood and adult obesity, are not distinguished in this table as only obesity is included in the benchmark dataset.

Table 11. The top 20 predicted microbes associated with *ASTHMA*

Ranking	Microbe	Evidence	Description
1	*Bifidobacterium*	PMID:30290688	Increased
2	*Helicobacter pylori*	NA	NA
3	*Haemophilus*	PMID:32072252	Increased
4	*Faecalibacterium*	PMID:26424567	Decreased
5	*Streptococcus*	PMID:25329665	Decreased
6	*Bifidobacterium adolescentis*	PMID:26840903	Decreased
7	*Akkermansia muciniphila*	PMID:35265071	Decreased
8	*Rothia*	PMID:26424567	Decreased
9	*Faecalibacterium prausnitzii*	PMID:30208875	Decreased
10	*Prevotella*	NA	NA
11	*Dialister*	PMID:36969260	Increased
12	*Bacteroidetes*	PMID:20052417	Decreased
13	*Veillonella dispar*	NA	NA
14	*Veillonella*	PMID:29445257	NA
15	*Pseudomonas*	PMID:25329665	Increased
16	*Neisseria*	PMID:37287344	Decreased
17	*Burkholderia*	PMID:39549985	Increased
18	*Pseudomonas aeruginosa*	NA	NA
19	*Roseburia*	PMID:29031597	NA
20	*Parascardovia*	NA	NA

Table 12. The top 20 potential microbes for *CROHN'S DISEASE*

Ranking	Microbe	Evidence	Description
1	*Actinobacteria*	NA	NA
2	*Prevotellaceae*	PMID:35890149	Increased
3	*Porphyromonas gingivalis*	NA	NA
4	*Alcaligenaceae*	NA	NA
5	*Methanobrevibacter smithii*	NA	NA
6	*Streptococcus sanguinis*	NA	NA
7	*Lactobacillus crispatus*	NA	NA
8	*Fusobacterium periodonticum*	PMID:37932491	Increased
9	*Tanerella forsythia*	NA	NA
10	*Streptococcus gordonii*	PMID:39438255	Increased
11	*Treponema denticola*	PMID:23060013	Increased
12	*Streptococcus oralis*	PMID:34646784	Increased
13	*Lactobacillus iners*	NA	NA
14	*Streptococcus constellatus*	PMID:34725610	NA
15	*Campylobacter rectus*	PMID:31522142	NA
16	*Eikenella corrodens*	PMID:29574823	Increased
17	*Selenomonas noxia*	NA	NA
18	*Aggregatibacter actinomycetemcomitans*	PMID:36768711	Increased
19	*Capnocytophaga sputigena*	NA	NA
20	*Capnocytophaga ochracea*	NA	NA

Conclusion

In this study, we develop a computational approach SABMDA based on ensemble learning for microbe-disease association prediction. Our method first fuses multiple information from both microbes and diseases as input features. We then develop two matrix completion strategies to recover unknown microbe-disease associations. We conduct comprehensive experiments, and results demonstrate the superiority of our model in inferring new associations between microbes and diseases.

The excellent performance of our method can be attributed to three factors. The first is that we use reliable biomedical information as benchmark datasets in this study. The second is that we integrate multiple information from both microbes and diseases as input features. The third factor is that we apply ensemble learning for association prediction. Combining two matrix completion algorithms improve the inference performance. Meanwhile, it should be noted that the mechanism of how microbes affecting human diseases is complex. Our method predicts only microbe-disease associations. These associations are not true causal relationships between microbes and diseases. Details about how microbes positively or negatively contribute to human health need to be further investigated. Revealing the real causal effects between them would provide more useful help for biomedical research, which is a future research direction.

Table 13. The top 20 potential microbe-disease associations predicted by SABMDA

Ranking	Microbe	Disease	Evidence	Description
1	*Veillonella*	Ankylosing spondylitis	PMID:30944880	Increased
2	*Lactobacillus*	Ankylosing spondylitis	PMID:36548483	Decreased
3	*Blautia*	Autoimmune hepatitis	PMID:32640728	Increased
4	*Veillonella*	Biliary atresia	PMID:34630385	Increased
5	*Actinomyces*	Obesity	NA	NA
6	*Coprococcus*	Ankylosing spondylitis	PMID:37875269	Increased
7	*Faecalibacterium*	Autoimmune hepatitis	PMID:37945156	Increased
8	*Parabacteroides*	Ulcerative colitis	PMID:36547911	Increased
9	*Faecalibacterium*	Alzheimer's disease	NA	NA
10	*Ruminococcaceae*	Major depressive disorder	PMID:32229219	Decreased
11	*Leptotrichia*	Crohn's disease	PMID:38849764	Increased
12	*Lactobacillus*	Pancreatitis	NA	NA
13	*Bifidobacterium*	Primary biliary cholangitis	PMID:36287108	Increased
14	*Rothia*	Colorectal cancer	PMID:33844851	Increased
15	*Parabacteroides*	Primary biliary cholangitis	NA	NA
16	*Veillonella*	Colorectal cancer	PMID:36539569	Increased
17	*Coprococcus*	Short bowel syndrome	NA	NA
18	*Parabacteroides*	Coronary artery disease	PMID:35343796	Decreased
19	*Bifidobacterium*	Osteoporosis	PMID:37118342	Decreased
20	*Faecalibacterium*	Osteoporosis	PMID:37810879	Decreased

Key Points

- We propose an ensemble learning method SABMDA to predict novel disease-associated microbes, in which two matrix completion strategies are developed and used for prediction.
- Combination of the two matrix completion strategies improves microbe-disease association prediction.
- Comprehensive experiments demonstrate SABMDA outperforms recent state-of-the-art methods significantly.

Supplementary data

Supplementary data is available at *Briefings in Bioinformatics* online.

Conflict of interest: The authors have declared that no competing interests exist.

Funding

This work was supported by Jiangxi Provincial Natural Science Foundation, China (20242BAB25083).

Data availability

The data and source code for this study are available at https://github.com/IamChenHailin/SABMDA.

References

1. Human Microbiome Project C. A framework for human microbiome research. *Nature* 2012;**486**:215–21. https://doi.org/10.1038/nature11209
2. Ley RE, Peterson DA, Gordon JI. Ecological and evolutionary forces shaping microbial diversity in the human intestine. *Cell* 2006;**124**:837–48. https://doi.org/10.1016/j.cell.2006.02.017
3. Eckburg PB, Bik EM, Bernstein CN. *et al.* Diversity of the human intestinal microbial flora. *Science* 2005;**308**:1635–8. https://doi.org/10.1126/science.1110591
4. Peterson J, Garges S, Giovanni M. *et al.* The NIH human microbiome project. *Genome Res* 2009;**19**:2317–23. https://doi.org/10.1101/gr.096651.109
5. Blaser MJ. Harnessing the power of the human microbiome. *Proc Natl Acad Sci* 2010;**107**:6125–6. https://doi.org/10.1073/pnas.1002112107
6. Nicholson JK, Holmes E, Kinross J. *et al.* Host-gut microbiota metabolic interactions. *Science* 2012;**336**:1262–7. https://doi.org/10.1126/science.1223813
7. Althani AA, Marei HE, Hamdi WS. *et al.* Human microbiome and its association with health and diseases. *J Cell Physiol* 2016;**231**:1688–94. https://doi.org/10.1002/jcp.25284
8. Wen Z, Yan C, Duan G. *et al.* A survey on predicting microbe-disease associations: biological data and computational methods. *Brief Bioinform* 2021;**22**:bbaa157. https://doi.org/10.1093/bib/bbaa157
9. Niu Y-W, Qu C-Q, Wang G-H. *et al.* RWHMDA: random walk on hypergraph for microbe-disease association prediction. *Front Microbiol* 2019;**10**:1578. https://doi.org/10.3389/fmicb.2019.01578
10. Yan C, Duan G, Wu F-X. *et al.* BRWMDA: predicting microbe-disease associations based on similarities and bi-random walk on disease and microbe networks. *IEEE/ACM Trans Comput Biol Bioinform* 2019;**17**:1595–604. https://doi.org/10.1109/TCBB.2019.2907626
11. Wang L, Wang Y, Li H. *et al.* A bidirectional label propagation based computational model for potential microbe-disease association prediction. *Front Microbiol* 2019;**10**:684. https://doi.org/10.3389/fmicb.2019.00684
12. Huang Y-A, You Z-H, Chen X. *et al.* Prediction of microbe–disease association from the integration of neighbor and graph with collaborative recommendation model. *J Transl Med* 2017;**15**:1–11. https://doi.org/10.1186/s12967-017-1304-7
13. Wang F, Huang Z-A, Chen X. *et al.* LRLSHMDA: Laplacian regularized least squares for human microbe–disease association prediction. *Sci Rep* 2017;**7**:7601. https://doi.org/10.1038/s41598-017-08127-2

14. Zou S, Zhang J, Zhang Z. Novel human microbe-disease associations inference based on network consistency projection. *Sci Rep* 2018;**8**:8034. https://doi.org/10.1038/s41598-018-26448-8
15. Qu J, Zhao Y, Yin J. Identification and analysis of human microbe-disease associations by matrix decomposition and label propagation. *Front Microbiol* 2019;**10**:291. https://doi.org/10.3389/fmicb.2019.00291
16. Wu C, Gao R, Zhang Y. mHMDA: human microbe-disease association prediction by matrix completion and multi-source information. *IEEE Access* 2019;**7**:106687–93. https://doi.org/10.1109/ACCESS.2019.2930453
17. Shi J-Y, Huang H, Zhang Y-N. *et al.* BMCMDA: a novel model for predicting human microbe-disease associations via binary matrix completion. *BMC Bioinformatics* 2018;**19**:85–92. https://doi.org/10.1186/s12859-018-2274-3
18. Li H, Wang Y, Zhang Z. *et al.* Identifying microbe-disease association based on a novel back-propagation neural network model. *IEEE/ACM Trans Comput Biol Bioinform* 2020;**18**:2502–13. https://doi.org/10.1109/TCBB.2020.2986459
19. Wang L, Wang Y, Xuan C. *et al.* Predicting potential microbe–disease associations based on multi-source features and deep learning. *Brief Bioinform* 2023;**24**:bbad255. https://doi.org/10.1093/bib/bbad255
20. Bennett J, Elkan C, Liu B. *et al.* Kdd cup and workshop 2007. *ACM SIGKDD Explorations Newsletter* 2007;**9**:51–2. https://doi.org/10.1145/1345448.1345459
21. Bertsekas DP. Nonlinear programming. *Journal of the Operational Research Society* 1997;**48**:334–4. https://doi.org/10.1057/palgrave.jors.2600425
22. Elman HC, Golub GH. Inexact and preconditioned Uzawa algorithms for saddle point problems. *SIAM Journal on Numerical Analysis* 1994;**31**:1645–61. https://doi.org/10.1137/0731085
23. Cai J-F, Candès EJ, Shen Z. A singular value thresholding algorithm for matrix completion. *SIAM Journal on Optimization* 2010;**20**:1956–82. https://doi.org/10.1137/080738970
24. Candes E, Recht B. Simple bounds for recovering low-complexity models. *Mathematical Programming* 2013;**141**:577–89. https://doi.org/10.1007/s10107-012-0540-0
25. Chen C, He B, Yuan X. Matrix completion via an alternating direction method. *IMA Journal of Numerical Analysis* 2012;**32**:227–45. https://doi.org/10.1093/imanum/drq039
26. Li C-N, Shao Y-H, Yin W. *et al.* Robust and sparse linear discriminant analysis via an alternating direction method of multipliers. *IEEE Transactions on Neural Networks and Learning Systems* 2019;**31**:915–26. https://doi.org/10.1109/TNNLS.2019.2910991
27. Li J-Q, Rong Z-H, Chen X. *et al.* MCMDA: matrix completion for MiRNA-disease association prediction. *Oncotarget* 2017;**8**:21187–99. https://doi.org/10.18632/oncotarget.15061
28. Yang M, Luo H, Li Y. *et al.* Drug repositioning based on bounded nuclear norm regularization. *Bioinformatics* 2019;**35**:i455–63. https://doi.org/10.1093/bioinformatics/btz331
29. Chen H, Chen K. Predicting disease-associated microbes based on similarity fusion and deep learning. *Brief Bioinform* 2024;**25**:bbae550. https://doi.org/10.1093/bib/bbae550
30. Ning Q, Zhao Y, Gao J. *et al.* AMHMDA: attention aware multi-view similarity networks and hypergraph learning for miRNA-disease associations identification. *Brief Bioinform* 2023;**24**:bbad094. https://doi.org/10.1093/bib/bbad094
31. Peng W, He Z, Dai W. *et al.* MHCLMDA: multihypergraph contrastive learning for miRNA-disease association prediction. *Brief Bioinform* 2023;**25**:bbad524. https://doi.org/10.1093/bib/bbad524
32. Liu H, Bing P, Zhang M. *et al.* MNNMDA: predicting human microbe-disease association via a method to minimize matrix nuclear norm. *Comput Struct Biotechnol J* 2023;**21**:1414–23. https://doi.org/10.1016/j.csbj.2022.12.053
33. Luo J, Long Y. NTSHMDA: prediction of human microbe-disease association based on random walk by integrating network topological similarity. *IEEE/ACM Trans Comput Biol Bioinform* 2020;**17**:1341–51. https://doi.org/10.1109/TCBB.2018.2883041
34. Huang Z, Shi J, Gao Y. *et al.* HMDD v3.0: a database for experimentally supported human microRNA-disease associations. *Nucleic Acids Res* 2019;**47**:D1013–7. https://doi.org/10.1093/nar/gky1010
35. Kopelman PG. Obesity as a medical problem. *Nature* 2000;**404**:635–43. https://doi.org/10.1038/35007508
36. Borish L. The immunology of asthma: asthma phenotypes and their implications for personalized treatment. *Ann Allergy Asthma Immunol* 2016;**117**:108–14. https://doi.org/10.1016/j.anai.2016.04.022
37. Baumgart DC, Sandborn WJ. Crohn's disease. *The Lancet* 2012;**380**:1590–605. https://doi.org/10.1016/S0140-6736(12)60026-9

Briefings in Bioinformatics, 2025, **26(2)**, bbaf075
https://doi.org/10.1093/bib/bbaf075
Problem Solving Protocol