frontiers | Frontiers in Microbiology

ORIGINAL RESEARCH
published: 31 May 2022
doi: 10.3389/fmicb.2022.919380

Metapath Aggregated Graph Neural Network and Tripartite Heterogeneous Networks for Microbe-Disease Prediction

*Yali Chen and Xiujuan Lei**

School of Computer Science, Shaanxi Normal University, Xi'an, China

More and more studies have shown that understanding microbe-disease associations cannot only reveal the pathogenesis of diseases, but also promote the diagnosis and prognosis of diseases. Because traditional medical experiments are time-consuming and expensive, many computational methods have been proposed in recent years to identify potential microbe-disease associations. In this study, we propose a method based on heterogeneous network and metapath aggregated graph neural network (MAGNN) to predict microbe-disease associations, called MATHNMDA. First, we introduce microbe-drug interactions, drug-disease associations, and microbe-disease associations to construct a microbe-drug-disease heterogeneous network. Then we take the heterogeneous network as input to MAGNN. Second, for each layer of MAGNN, we carry out intra-metapath aggregation with a multi-head attention mechanism to learn the structural and semantic information embedded in the target node context, the metapath-based neighbor nodes, and the context between them, by encoding the metapath instances under the metapath definition mode. We then use inter-metapath aggregation with an attention mechanism to combine the semantic information of all different metapaths. Third, we can get the final embedding of microbe nodes and disease nodes based on the output of the last layer in the MAGNN. Finally, we predict potential microbe-disease associations by reconstructing the microbe-disease association matrix. In addition, we evaluated the performance of MATHNMDA by comparing it with that of its variants, some state-of-the-art methods, and different datasets. The results suggest that MATHNMDA is an effective prediction method. The case studies on asthma, inflammatory bowel disease (IBD), and coronavirus disease 2019 (COVID-19) further validate the effectiveness of MATHNMDA.

Keywords: microbe-disease associations, heterogeneous network, metapath aggregated graph neural network, multi-head attention mechanism, COVID-19

OPEN ACCESS

Edited by:
Qi Zhao,
University of Science and Technology Liaoning, China

Reviewed by:
Wei Peng,
Kunming University of Science and Technology, China
Xing Chen,
China University of Mining and Technology, China

***Correspondence:**
Xiujuan Lei
xjlei@snnu.edu.cn

Specialty section:
This article was submitted to Systems Microbiology, a section of the journal Frontiers in Microbiology

Received: *13 April 2022*
Accepted: *29 April 2022*
Published: *31 May 2022*

Citation:
Chen Y and Lei X (2022) Metapath Aggregated Graph Neural Network and Tripartite Heterogeneous Networks for Microbe-Disease Prediction. Front. Microbiol. 13:919380. doi: 10.3389/fmicb.2022.919380

INTRODUCTION

The microorganisms related to the human body mainly include eukaryotes, archaea, bacteria, fungi, and viruses [Human Microbiome Project (HMP), 2012]. These microorganisms form different microbial communities and parasitize in different parts of the human body, such as the skin, mouth, genitalia, intestinal tract, and other parts. Studies have shown that the number of microbes in the

adult intestine is equivalent to 10 times that of human cells (Sender et al., 2016), which indicates that the microbial community in the human body is relatively large. Microbes are generally beneficial to the human body. For example, by fermenting food ingredients that cannot be digested by the host, gut microbes can promote nutrient and energy absorption (Gill et al., 2006; Marco et al., 2017). The *Bifidobacteria* in the human intestine can produce lactic acid and acetic acid after fermentation, which can promote the absorption of iron and vitamin D. Therefore, a set of balanced microbes can keep the human body away from physiological disorders, but the imbalance or decline of the microbial community can harm the human host and cause diseases. For example, a study has found that compared to normal children, children with asthma would have a smaller number of *Faecalibacterium*, *Lachnospira*, *Veillonella*, and *Rothia* (Arrieta et al., 2015). Another study found that the relative abundance of *Enterococcus*, *Escherichia*/*Shigella*, *Klebsiella*, *Streptococcus,* and *Peptostreptococcus* in the intestinal flora of patients with colorectal cancer was increased (Wang et al., 2012). These studies have shown that identifying the relationship between microbes and diseases can help us understand the pathogenesis of the disease, so as to carry out more targeted treatment. Therefore, determining the relationship between microbes and diseases has become a key research topic in the current bioinformatics field.

Verifying the relationship between microbes and diseases through biological experiments is a time-consuming and expensive task. Therefore, many computational models have been proposed to predict the association between microbes and diseases. Wang et al. (2021) wrote a review on circular RNAs and complex diseases, which classified the prediction models of circRNA-disease associations. Inspired by this study, we can divide these computational models into four types according to the differences in the microbe-disease association prediction strategies based on heterogeneous networks: path-based methods, random walk methods, bipartite local models, and matrix decomposition methods (Wen et al., 2021). Path-based methods are widely used in association prediction (Zhang et al., 2021; Liu et al., 2022a). They make predictions by calculating path-based scores between microbe nodes and disease nodes. For example, Chen X. et al. (2016) proposed the first model KATZHMDA to predict microbe-disease associations, which calculated the predicted probability score according to the walking step length and walking times between the two nodes in the microbe-disease network. Huang et al. (2017) proposed a computational model PBHMDA based on the depth-first search to predict potential microbes associated with diseases. Fan et al. (2018) developed a new model MDPH_HMDA to predict microbe-disease associations by integrating multi-source data and path-based HeteSim score. The random walk has aroused extensive interest in the field of microbe-disease prediction. For instance, Yan et al. (2019) proposed a prediction model BRWMDA based on similarity and bi-random walk to predict potential microbe and disease associations. Luo and Long (2018) proposed a computational model NTSHMDA based on random walk and network topology similarity to predict the associations between microbes and diseases. Wu et al. (2018) developed a method named PRWHMDA, which attempted to infer potential microbe-disease pairs by random walk on the heterogeneous network with Particle Swarm Optimization (PSO). Bipartite local models are also common methods, which work independently on the basis of both sides of a microbe-disease pair and can be combined to yield a definitive prediction result. For example, Zou et al. (2018) proposed a method called NCPHMDA that utilized the network consistency projection to predict microbe-disease associations. Wang et al. (2017) constructed a semi-supervised computational model LRLSHMDA based on a Laplacian regularization least squares classifier to predict the associations between microbes and diseases. In addition, some prediction models for microbe-disease associations were developed based on matrix factorization techniques. For instance, He et al. (2018) presented a method called GRNMFHMDA, which incorporated weighted K-nearest known neighbors to predict microbe-disease associations. Shen et al. (2017) developed a computational model of CMFHMDA, which used collaborative matrix factorization to reconstruct the association matrices between diseases and microbes. Wang Y. et al. (2022) proposed a method HNGFL based on heterogeneous network and global graph feature learning to predict microbe-disease association. In addition to these computational models, several review articles on microbe-disease associations have been published. For example, Pan et al. (2022) developed a comprehensive approach to predict associations between genomics, proteinomics, transcriptomics, microbiome, metabolomics, pathomics, radiomics, drug, symptoms, environment factors, and disease networks. Wang L. et al. (2022) provided a comprehensive review on predicting pairwise relationships between human microbes, drugs, and diseases, from biological data to computational models. Wen et al. (2021) provided a survey on predicting microbe-disease associations based on biological data and computational methods.

Although the above-mentioned methods have achieved relatively stable prediction performance in the association prediction task of microbes and diseases, there are still some limitations and deficiencies. First, the vast majority of methods make predictions based on small-scale datasets, which makes them unable to obtain accurate predictions when it comes to new diseases (or new microbes) due to a lack of training data. Second, microbe imbalance (or the occurrence of disease) is not influenced by a single factor. Some studies have shown that microbes participate in drug absorption and metabolism, thereby regulating drug efficacy and drug toxicity for disease (Zimmermann et al., 2021). However, the above-mentioned methods are only based on microbes and diseases, which makes these models unable to obtain accurate prediction results due to the lack of more semantic information about microbes and diseases in the prediction process.

Therefore, with the discovery of multivariate biological data, the heterogeneous graph embedding method is increasingly applied to relational prediction. It can learn semantic and structural information between nodes to compensate for the poor prediction performance due to the small amount of known associated data. For example, Lei and Wang (2020)

proposed a method based on Node2vec and a heterogeneous network scoring mechanism, called LGRSH, to predict the association between microbes and diseases. Liu et al. (2022b) proposed a method to identify miRNA-disease associations via deep forest ensemble learning based on autoencoder. Yang et al. (2022) proposed a DeepWalk-based method to predict lncRNA-miRNA associations via a lncRNA-miRNA-disease-protein-drug graph. Zhu et al. (2018) proposed a method using Metapath2vec to predict drug-gene interactions. Lei et al. (2021) developed a method, called CDWBMS, to predict circRNA-disease associations based on an improved weighted biased meta-structure. Zhang et al. (2020) adopted metapath2vec++ and matrix factorization to predict circRNA-disease associations. All the heterogeneous graph embedding methods have some limitations when applied to association prediction, such as ignoring the information of multiple nodes, discarding all intermediate nodes on the metapath, or only using a single metapath. This will affect the predictive performance of the model.

To deal with the above-mentioned issues, we developed a novel method based on a metapath aggregated graph neural network (MAGNN) and tripartite heterogeneous network for microbe-disease association prediction named MATHNMDA. In particular, we integrate information from different sources, such as microbe-disease associations, microbe-drug interactions, and disease-drug interactions, to construct a tripartite heterogeneous network of microbe-drug-disease. Further, we feed the heterogeneous network to MAGNN. For each layer of MAGNN, we first use intra-metapath aggregation with a multi-head attention mechanism to extract the structural and semantic information of the metapath instance. After that, we further apply inter-metapath aggregation with an attention mechanism to fuse latent vectors of multiple metapaths. Finally, we take the output of the MAGNN as the final embedding features of the microbe node and disease node, and make predictions. In order to verify the predictive performance of MATHNMDA, we carried out cross-validation experiments, and the results indicate that MATHNMDA can effectively identify potential disease-related microbes.

Overall, our main contributions are as follows:

(1) We expand known microbe-disease association data by integrating multiple databases, and construct a tripartite heterogeneous network by introducing drug-disease associations and microbe-drug associations. We further apply MAGNN to predict microbe-disease associations.
(2) We use intra-metapath aggregation with the multi-head attention mechanism to learn the topological information and semantic information embedded in the internal nodes of metapath, so that the embedding learned by the target node is more comprehensive.
(3) We use inter-metapath aggregation with an attention mechanism to aggregate the embeddings of different metapaths for target nodes (microbe nodes or disease nodes).
(4) We conduct a case study of coronavirus disease 2019 (COVID-19) to verify the effectiveness of the MATHNMDA model.

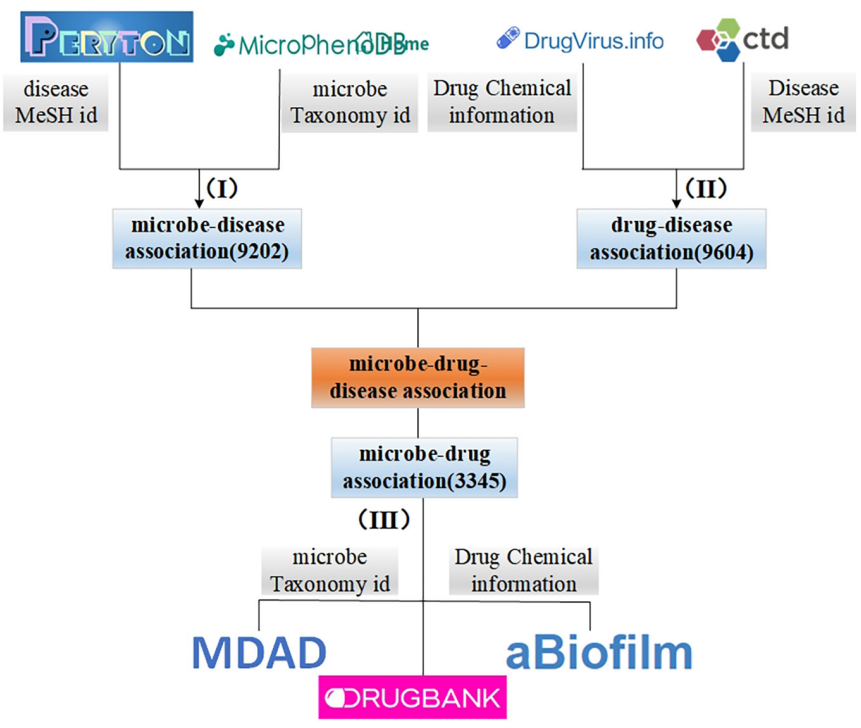

FIGURE 1 | The processing and integration process of microbe-drug-disease association data.

MATERIALS AND METHODS

Dataset

In this study, we integrate the information obtained from different sources. First, we collect microbe and disease association data from Peryton (Skoufos et al., 2020) and MicroPhenoDB (Yao et al., 2021). Among them, Peryton includes more than 7,900 relationships between 43 diseases and 1,396 microbes. The data in MicroPhenoDB are collected from some public datasets, such as Human Microbe-Disease Association Database (HMDAD; Kong et al., 2017), Disbiome (Yorick et al., 2018), Virulence Factor Database (VFDB; Chen L. et al., 2016), etc. MicroPhenoDB has 5,565 relationships between 515 diseases and 1,717 microbes. After eliminating redundancy for the same diseases and microbes, we obtain a total of 9,202 associations between 538 diseases and 2,491 microbes. Furthermore, we collect data about microbes and their related drugs from Microbe-Drug Association Database (MDAD; Sun et al., 2018), drugVirus (Andersen et al., 2020), and aBiofilm (Akanksha et al., 2017), and remove redundant records to obtain a total of 132 microbes and 1,933 drugs and 3,345 microbe-drug associations. Then, we download disease-drug interaction data from drugBank (Wishart et al., 2017) and Comparative Toxicogenomics Database (CTD; Davis et al., 2012) databases, and we obtain 9,604 interactions between 127 diseases and 247 drugs after de-redundancy. **Figure 1** illustrates the integration process of microbe-drug data, drug-disease data, and microbe-disease data. It is worth noting that in this study, we unified the disease, microbe, and drug according to the MESH id of the disease, the taxonomy id of the microbe and the chemical information of the drug, disease-related drugs, are included in drugs related to microbes.

Construction of Microbe-Drug-Disease Tripartite Heterogeneous Network

In this study, we use microbe-disease, microbe-drug, and disease-drug associations to build a tripartite network. The relationship between microbes, drugs, and diseases is shown in **Figure 2A**. A certain microbial imbalance can lead to certain diseases. and the pathogenesis of a certain disease will be affected by certain microbial communities. Some drugs can treat some diseases, and certain diseases can be treated with certain drugs. Microbes can regulate the activity and toxicity of drugs (Zimmermann M. et al., 2019). Drugs in turn can change the diversity and function of microbial communities. Suppose M, C, and D, respectively, represent all the sets of microbes, drugs, and diseases in the network, $m_i \in M$ represents a microbe, $i = 1, 2, 3..., n_m$; $c_j \in C$ represents a drug, $j = 1, 2, 3..., n_c$; and $d_k \in D$ represents a disease, $k = 1, 2, 3..., n_d$. Construct a tripartite heterogeneous network based on the relationship among microbes, drugs, and diseases. Here, we can simplify it to an undirected and unweighted network to represent the existence of associations, as shown in **Figure 2B**. We further construct the microbe-disease adjacency matrix $B \in R^{n_m \times n_d}$, where n_m represents the number of microbes and n_d represents the number of diseases. If there is a known association between a microbe node i and a disease node j, the value of $B(i,j)$ is 1, otherwise, it is 0.

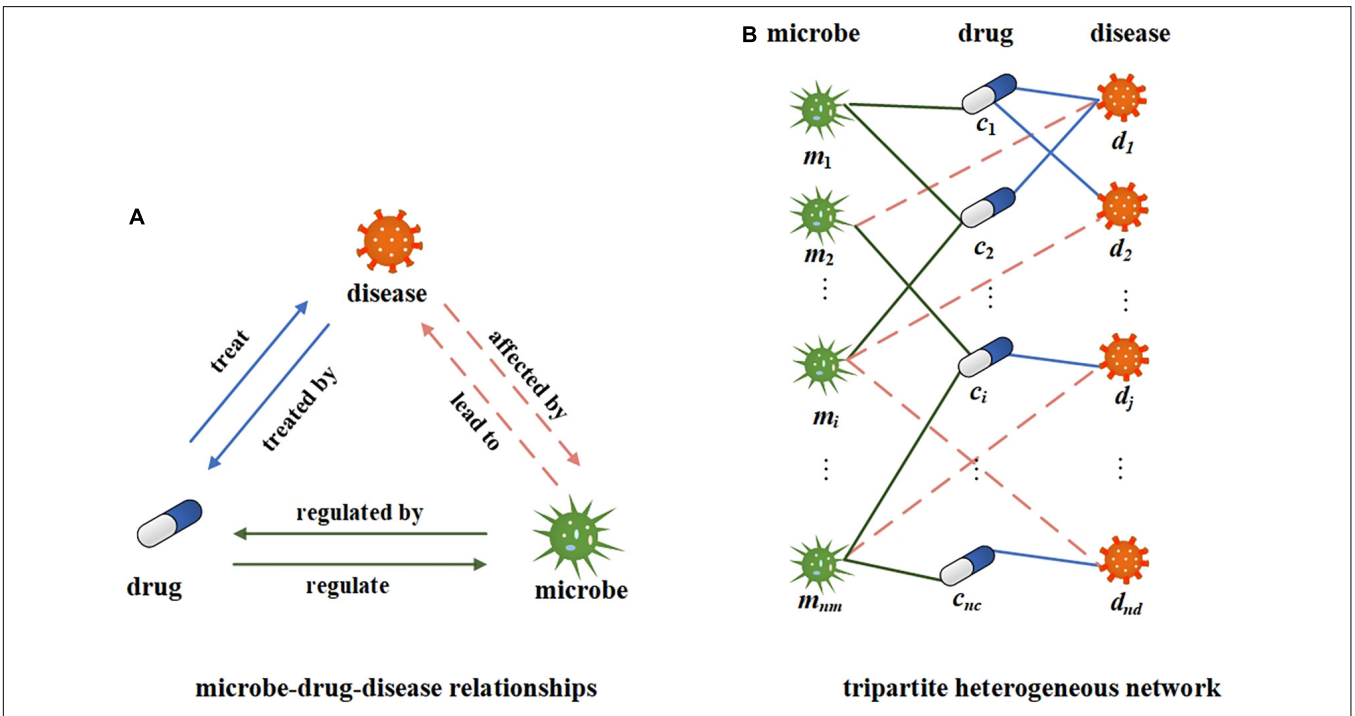

FIGURE 2 | Illustration of microbe-drug-disease relationship and heterogeneous networks. **(A)** The illustration of microbe-drug-diease relationships. **(B)** A tripartite network corresponding to the microbe-drug-disease relationships.

MATHNMDA

Our proposed MATHNMDA model consists of three main steps, as shown in **Figure 3**. The model takes heterogeneous microbe-drug-disease interaction network and MAGNN to predict microbe-disease associations. First, we take heterogeneous network as input of the MAGNN. Second, for each layer of the MAGNN, we use intra-metapath aggregation to learn the structural and semantic information embedded in the target node, metapath-based neighbor nodes, and the context between them. Third, we apply inter-metapath aggregation to combine the semantic information of all different metapaths. Finally, we take the output of the MAGNN as vector representations of microbe nodes and disease nodes, which can be used to predict potential microbe-disease associations.

Intra-Metapath Aggregation

In this study, we predict novel microbe-disease associations on the heterogeneous microbe-drug-disease interaction networks based on MAGNN. Given a heterogeneous network $G = (V, E)$, where V and E represent sets of nodes and edges, respectively, and the mapping functions of nodes and edges are δ: $V \rightarrow A$ and ψ: $E \rightarrow R$, A represents node types, R denotes edge types, and $|A| + |R| > 2$. Given a metapath M on the heterogeneous network G, we can define it as a path of the form $A_1 \rightarrow A_2 \rightarrow \ldots \rightarrow A_{n-1} \rightarrow A_n$, which can be abbreviated as $A_1A_2 \ldots A_{n-1}A_n$. The relationship between node types A_1 and A_n is $R = R_1 \circ R_2 \circ \ldots \circ R_{n-1}$, where $\circ$ represents the composite operation. That is to say, the relationship R is obtained by compositing the n–1 relationships of these $R_1, R_2, \ldots, R_{n-1}$. Therefore, a metapath can capture specific semantic information in the graph, and different metapaths represent different semantic information. For example, for the metapath microbe-drug-disease (abbreviated as *m-c-d*), drug *c* can act on microbe *m*, and drug *c* can be used to treat disease *d*, so microbe *m* may be associated with disease *d*. The key idea of intra-metapath aggregation is to learn structural and semantic information embedded in target nodes, metapath-based neighbors, and the context between them by encoding metapath instances under a certain metapath. Next, we introduce the process of intra-metapath aggregation in detail.

Given a metapath M, we define a sequence of nodes in G that follow the pattern of M as a metapath instance, defined as $M(i,j)$, which is represented as a metapath instance connecting the target node i and its neighbor node j based on the metapath. Here, $j \in N_i^M$, N_i^M represents the set of nodes connected to node i through the metapath instance $M(i,j)$. It is worth noting that if the metapath instance $M(i,j)$ is symmetric, $j \in N_i^M$ also includes node i itself. Then we define the intermediate node set of $M(i,j)$ as $TH(i,j)$, $TH(i,j) = M(i,j) / \{j, i\}$, where $\{j, i\}$ represents the set with elements i, j.

As mentioned before, intra-metapath aggregation learns structural and semantic information of target nodes by encoding metapath instances. Sun et al. (2019) proposed a method for knowledge graph embedding based on relational rotation in complex space, called RotatE. RotatE can model all relational patterns, so we use RotatE as the metapath instance encoder in this study. Given a metapath instance $M(i,j) = (i, th_2 \ldots, th_{n-1}, j)$, for convenience, let set $i = th_1$ and $j = th_n$. R_i represents the relationship between node th_i and node th_{i+1}, and the relationship vector is r_i. Therefore, for the metapath instance $M(i,j)$, RotatE can be defined as follows:

$$\begin{aligned} \Theta_1 &= \tilde{h}_{th_1} = \tilde{h}_i \\ \Theta_i &= \tilde{h}_{th_i} + \theta_{i-1} \odot r_i \\ h_{M(i,j)} &= \frac{\theta_n}{n} \end{aligned} \tag{1}$$

where $\tilde{h}_{th_i}$ and r_i are vectors of complex space, $\odot$ represents hadamard product, $hM(i,j) \in R^{d'}$, and d' is the dimension of $h_{M(i,j)}$. In which case, we get vector representation of the metapath instance $M(i,j)$. It is important to note that there may be multiple instances of the metapath M connecting nodes i and j, but we use $M(i,j)$ to represent a single instance here.

Graph attention network (GAT) is an effective graph representation learning tool, which represents the importance of neighbor nodes to the target node by assigning different weights to different neighbor nodes (Bian et al., 2021). Here, for target node i and metapath M related to i, we first use GAT to assign weights (attention coefficients) to metapath instances in M, thereby learning the importance of different metapath instances to target nodes. Then the features of different metapath instances are aggregated according to the obtained attention coefficients, which are represented as the feature vector of the target node i. Given a metapath instance $M(i,j)$, its attention coefficient can be defined as:

$$e_{ij}^M = LeakyReLU\left(\delta_M^T \left[\tilde{h}_i \parallel h_{M(i,j)}\right]\right) \tag{2}$$

where δ_M^T is the attention parameter of the metapath M, and $\parallel$ represents connection operation. To make the attention coefficients of different metapath instances comparable, we use the softmax function to normalize e_{ij}^M:

$$\alpha_{ij}^M = softmax\left(e_{ij}^M\right) = \frac{exp\left(e_{ij}^M\right)}{\sum_{k \in N_i^M} \exp\left(e_{ij}^M\right)} \tag{3}$$

Then, we aggregate the feature vectors of all metapath instances according to the activation function $\sigma(\cdot)$ to obtain the vector representation of node i based on the metapath M:

$$\alpha_{ij}^M = \sigma\left(\sum_{j \in N_i^M} \alpha_{ij}^M \cdot h_{M(i,j)}\right) \tag{4}$$

In this study, we further introduce a multi-head attention mechanism to stabilize the learning process of attention coefficients and reduce the influence of a single attention. Specifically, we independently repeat the attention mechanism K times and concatenate vector representation learned by each attention head. Therefore, the vector representation of node i can be further rewritten as follows:

$$h_i^M = \mathop{\Big\Vert}_{k=1}^{K} \sigma\left(\sum_{j \in N_i^M} \alpha_{ij}^M \cdot h_{M(i,j)}\right) \tag{5}$$

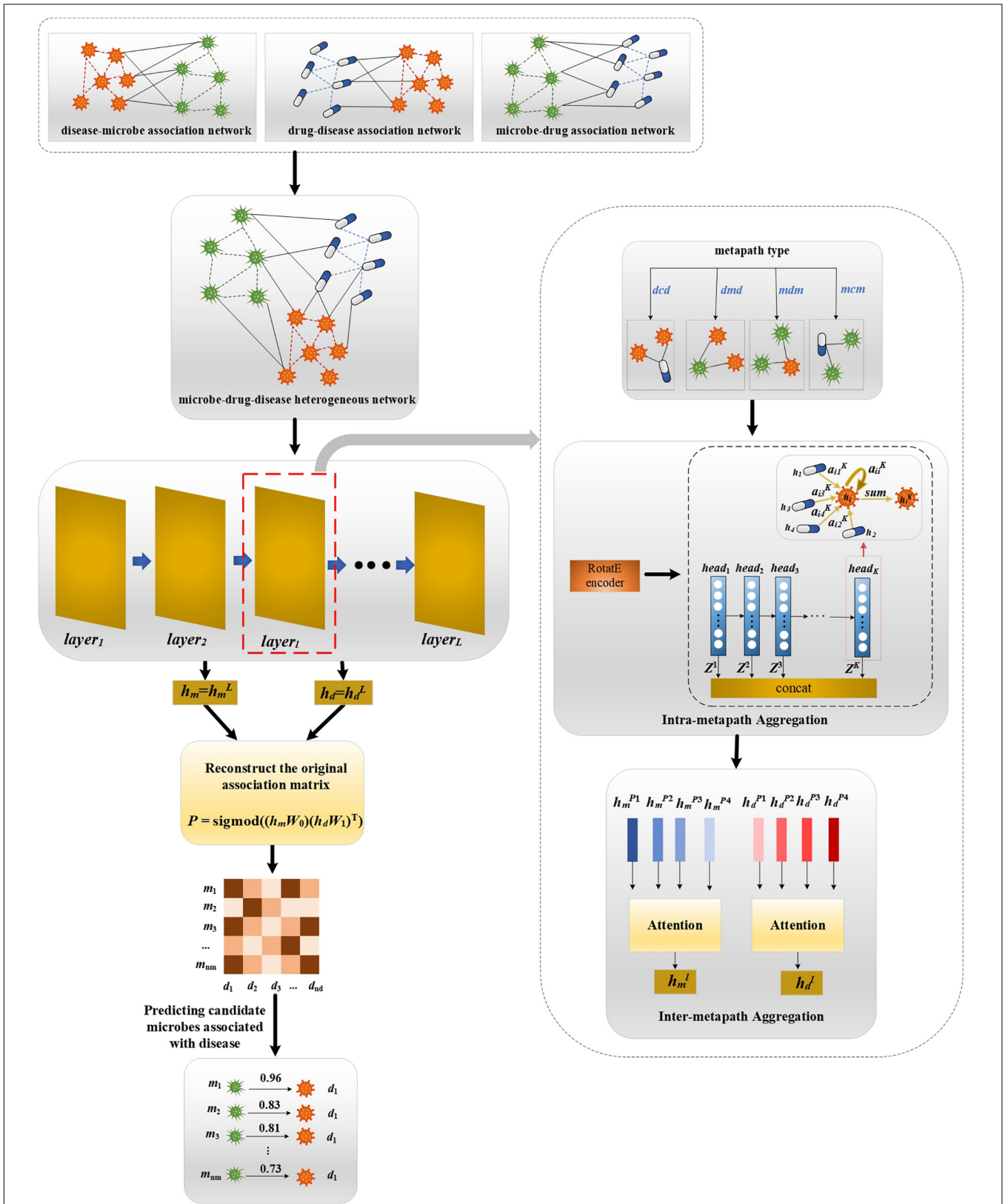

FIGURE 3 | The framework of MATHNMDA. *mcm*, *mdm*, *dmd,* and *dcd* are four metapaths, where *m* represents microbe, *c* represents drug, and *d* represents disease.

Since the metapath is undirected, each node in the metapath can be either a start node or an end node. Therefore, for the metapath set starting or ending with node type $a \in A$, it is denoted as $M_a = \{M_1, M_2, \cdots, M_S\}$, and S represents the number of metapaths. Intra-metapath aggregation obtains M metapath-specific vector representations of the target node $i \in V_a$, defined as $\left\{h_i^{M_1}, h_i^{M_2}, \cdots, h_i^{M_S}\right\}$.

Inter-Metapath Aggregation

After the intra-metapath aggregation of metapaths, we obtain the vector representation of a single metapath M for target node i. Then, we need to aggregate the semantic information and structural information of node i based on all metapaths of M_a, where S represents the number of metapaths. The node embedding set corresponding to these S metapaths is $\left\{h_i^{M_1}, h_i^{M_2}, \cdots, h_i^{M_S}\right\}$. A simple aggregation method between metapaths is to take the average of these node embeddings. However, because the importance of metapaths to node i in a heterogeneous network is different, we allocate weight for each metapath pattern through the attention mechanism, and then perform aggregation.

Specifically, given a metapath M_p, $M_p \in V_a$, we first transform these metapath-specific node vectors for all nodes $i \in V_a$ with the tanh function, and then take average value as feature of M_p:

$$S_{M_p} = \frac{1}{|V_a|} \sum_{i \in V_a} tanh\left(W_a \cdot h_i^{M_p} + b_a\right) \tag{6}$$

where W_a is the weight matrix of nonlinear transformation specific to node type a, and B_a is the corresponding bias, both of which are learnable parameters. V_a indicates all nodes of type a in the network.

Then we use the attention mechanism to calculate the importance of each metapath pattern for the target node i, and normalize the obtained attention coefficients by the softmax function. Then we fuse the corresponding vector representations of these metapaths to get the output of the target node i, as shown in Equation 7:

$$\begin{aligned} e_{M_p} &= c_a^T \cdot S_{M_p} \\ \beta_{M_p} &= softmax\left(e_{M_p}\right) = \frac{exp\left(e_{M_p}\right)}{\sum_{M_p \in M_a} \exp\left(e_{M_p}\right)} \\ h_i^{M_a} &= \sum_{M_p \in M_a} \beta_{M_p} \cdot h_i^{M_p} \end{aligned} \tag{7}$$

where c_a^T denotes the attention parameter, β_{M_P} denotes the normalized attention score, and M_P denotes the Pth metapath in M_a. $h_i^{M_a}$ represents the embedding vector of node i based on aggregation between metapaths.

MAGNN

The goal of a graph neural network (GNN) is to learn the low-dimensional vector representation of each node, which can be used for many downstream tasks, such as node clustering, node classification, and link prediction. Thus, we further apply an L-layer GNN to learn the low-dimensional representation vectors of microbe nodes and disease nodes. At each layer of the GNN, we use intra-metapath aggregation and inter-metapath aggregation to obtain vector representations of node-based metapath. In this way, we can define the low-dimensional representation for node i at the lth layer:

$$h_i^l = \sigma\left(W_o^l \cdot \left[h_i^{M_a}\right]^l\right) \tag{8}$$

where $\sigma(\cdot)$ is an activation function and W_o^l represents the weight vector at the lth layer. h_i^l represents the vector representation for node i at the lth layer, which is also the input of the $(l+1)$th layer. We define $h_i^0 = W_a \cdot X_i^a$, where W_a represents the linear transformation matrix of node type a and X_i is the original feature vector for node of type a. Here, we use one-hot encoding to initialize each type of node in the heterogeneous network.

Finally, we use vector representation of node i at the Lth layer to serve as the final embedding for nodes i:

$$h_i = h_i^L \tag{9}$$

where h_i^L represents vector representation of node i at the Lth layer.

Reconstruction of Microbe-Disease Association

After we get the final embeddings of all microbe nodes and disease nodes, we can predict new microbe-disease associations by reconstructing microbe-disease associations. Here we perform a simple inner product operation on the microbe and disease embeddings. In this case, each microbe-disease pair will receive a new score. Specifically, given a microbe node m and a disease node d, the predicted score C_{md} between them can be calculated as:

$$C_{md} = sigmoid\left(h_m^T \cdot h_d\right) \tag{10}$$

where h_m and h_d represent the final embeddings of microbes and diseases, respectively.

Optimization

Since our task is to predict microbe-disease associations, this is equivalent to a binary classification problem. So, here we use the cross-entropy function as the loss function and optimize through negative sampling:

$$L = -\sum_{(m,d) \in \mu} log\left(C_{md}\right) - \sum_{(m,d) \in \mu^-} log\left(-C_{md}\right) \tag{11}$$

where μ represents the set of positive samples, and μ^- represents the set of negative samples obtained by negative sampling.

RESULTS

In this section, we evaluate the performance of MATMNMDA through some experiments and analysis of the results. At the same time, we also analyze and adjust some parameters of the model in order to make better predictions.

Evaluation Metrics

In this study, we mainly use two metrics to evaluate the performance of the model, area under the receiver operating characteristic curve (AUC) and area under the precision–recall curve (AUPR), which are widely used in association prediction tasks.

AUC: This corresponds to the area of a planar graph bounded by the receiver operating characteristic (ROC) curve and horizontal axis, which can estimate the performance of binary classification models. The value of AUC is between 0 and 1. When it is closer to 1, the model performs better. In practical application, the advantages and disadvantages of different models can be compared by comparing the AUC values of different classification models.

AUPR: The precision–recall (PR) curve is also used to evaluate the classification ability of the model. In particular, the PR curve can collect more information when dealing with some imbalanced datasets. The area enclosed by the PR curve and the abscissa axis is called AUPR.

Baselines

In order to test the effectiveness of the MATMNMDA model, we compare it with six state-of-the-art methods based on the data processed in this study. Here, we calculate the AUC and AUPR values of these methods under the same conditions and analyze the results. The six baselines are as follows:

BRWMDA (Yan et al., 2019): It is a similarity-based and modified bi-random walk to predict associations between microbes and diseases.

KATZHMDA (Chen X. et al., 2016): It is a method to predict microbe-disease associations based on the katz metric.

LRLSHMDA (Wang et al., 2017): It is a semi-supervised model to predict microbe-disease associations by introducing a Gaussian kernel and Laplacian regularization.

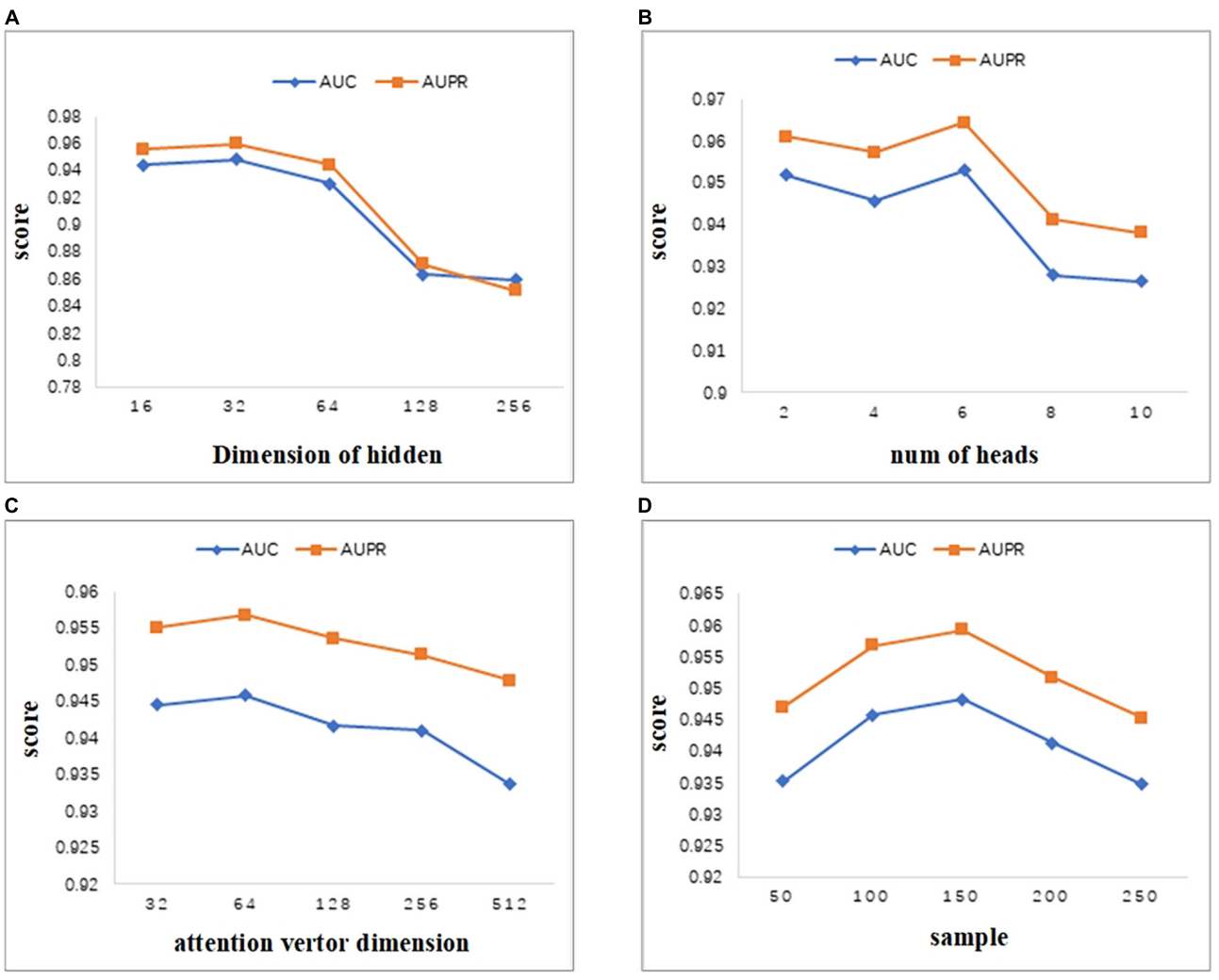

FIGURE 4 | Parameter analysis. **(A)** Comparison of AUC and AUPR for different hidden layer dimensions. **(B)** Comparison of AUC and AUPR of attention heads for different multi-attention mechanisms. **(C)** Comparison of AUC and AUPR for different attention vector dimensions. **(D)** Comparison of AUC and AUPR for different numbers of neighbors sampled by nodes.

NCPHMDA (Zou et al., 2018): It uses network consistent projections to predict microbe-disease associations.

NTSHMDA (Luo and Long, 2018): Predicting microbe-disease associations using heterogeneous network topological similarity and random walks.

CRPGCN (Ma et al., 2021): It is a method based on graph convolutional network (GCN) and random walk with restart (RWR), which was proposed for the cirRNA-disease association prediction task. Here, we use it as a baseline method for the prediction of microbe-disease association.

We compare MATMNMDA with these six baseline methods under the same conditions. For the CRPGCN method, the similarity of microbes and diseases is calculated in the same way as BRWMDA. For the MATMNMDA model, we first perform negative sampling on the microbe-drug-disease heterogeneous network. The positive and negative sample ratios of the training set, validation set, and test set are 1:1, and the proportion of the training set, validation set, and test set is 8:1:1, respectively. We randomly initialize vector representations of microbe nodes, drug nodes, and disease nodes. The Adam optimizer is used to optimize the model. The dropout and early stopping mechanisms are used to prevent overfitting. Here, according to the extensive literature (Phaisangittisagul, 2016), we set the value of dropout to 0.5. We train the model 100 times.

Parameter Analysis

In this section, we analyze the sensitivity of parameters. As we all know, important parameters will affect the performance of the model, so it is very necessary to conduct parameter analysis for the model. Some important parameters involved in the MATMNMDA model include the dimension of hidden layer, number of heads in the multi-head attention mechanism, dimension of attention vector, and number of neighbors sampled by the nodes in the experiment. We analyze these four parameters in turn and evaluate their impact on model performance.

As can be seen from **Figure 4A**, we set the dimension of the hidden layer to 16, 32, 64, 128, 256. As the dimension of the hidden layer increases, the performance of the model first increases. When the dimension reaches 32, both AUC and AUPR reach the maximum value. As the dimension continues to increase, the performance of the model begins to decrease gradually. Therefore, in this study, we set the embedding dimension of the hidden layer as 32. When the dimension changes between 16 and 256, the values of AUC and AUPR vary greatly. Thus, the MATMNMDA model is sensitive to the dimension of the hidden layer.

MATMNMDA model adopts a multi-attention mechanism to stabilize the process of attention coefficient learning. **Figure 4B** shows the influence of the number of attention heads in the multi-attention mechanism on model performance. We change the number of attention heads from 2 to 10 by step 2. It can be seen that when the number of attention heads is set to 6, the model has the best performance. **Figure 4C** shows the influence of the dimension of the attention vector. The dimension of the attention vector changes between 32 and 512. It can be observed that the vector dimension is too small or too large, which is not good for the performance of the model. Specifically, if the dimension of the attention vector is too large, it may lead to overfitting, which will degrade the performance of the model. When the dimension is set to 64, we can obtain better prediction ability.

In the MATMNMDA model, intra-metapath aggregation involves aggregating features of neighbor nodes to represent the representation of the current target node. Therefore, we analyze the number of neighbor nodes. In **Figure 4D**, the number of neighbor nodes is selected from {50,100,150,200,250}. It can be seen that when the number of neighbor nodes is too small or too large, the performance of the model is not very good. Specifically, if the number of neighbor nodes is too small, the structural information and semantic information of the target node may not be so comprehensive, while too large may cause noise. Therefore, we set the number of neighbor nodes to 150.

Ablation Study

As mentioned in the Introduction section, the previous heterogeneous network embedding methods have the following problems: (1) They only consider the neighbors based on the metapath, and do not consider the intermediate nodes inside the metapath. (2) In the metapath-based embedding, only the single best metapath is considered, and our model is proposed based on these problems. Therefore, in order to verify the effectiveness of each module of our model, we further conduct experiments on different variants of the MATMNMDA model. Taking MATMNMDA as a reference model, here we tested three variants of it.

MATMNMDA_nb: It only considers metapath-based neighbor nodes and does not consider intermediate nodes.

MATMNMDA_sm: It only considers the single best metapath.

MATMNMDA_avg: It replaces the RotatE with a mean encoder.

Figures 5, **6** show the comparison results of the MATMNMDA model and its variants. We can see that the MATMNMDA model has the highest AUC and AUPR. Followed by MATMNMDA_avg, MATMNMDA_sm has the worst performance. Comparing MATMNMDA and MATMNMDA_avg, we find that the MATMNMDA model performs better, which is because the mean encoder essentially treats metapath instances as a set and ignores the information embedded in the sequential structure of metapaths, while RotatE can be modeled according to the sequential structure of metapaths, thereby preserving the information embedded in the sequential structure of metapaths, so RotatE helps to improve the performance of the model by a small amount. Comparing MATMNMDA and MATMNMDA_nb, we can find that considering the intermediate nodes inside the metapath can help the model to obtain more structural information and thus improve the performance of the model. The results of MATMNMDA and MATMNMDA_sm show that the model performance can be significantly improved by combining multiple metapaths.

Comparison With Baselines

We run these baseline methods with default parameters. **Figures 7**, **8** show the performance of different methods. Our model achieves the highest prediction results on these two

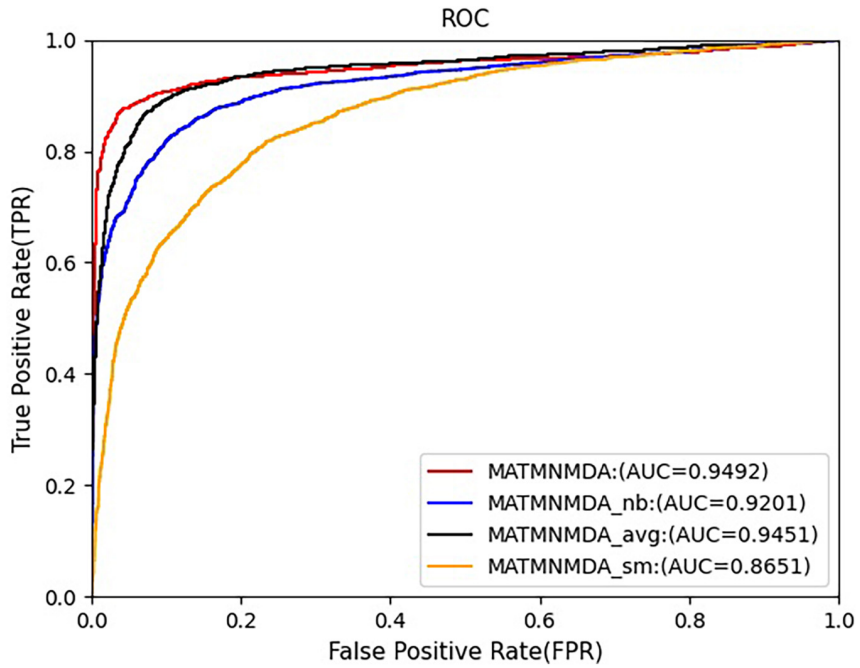

FIGURE 5 | Comparison of AUC for MATMNMDA and its variants.

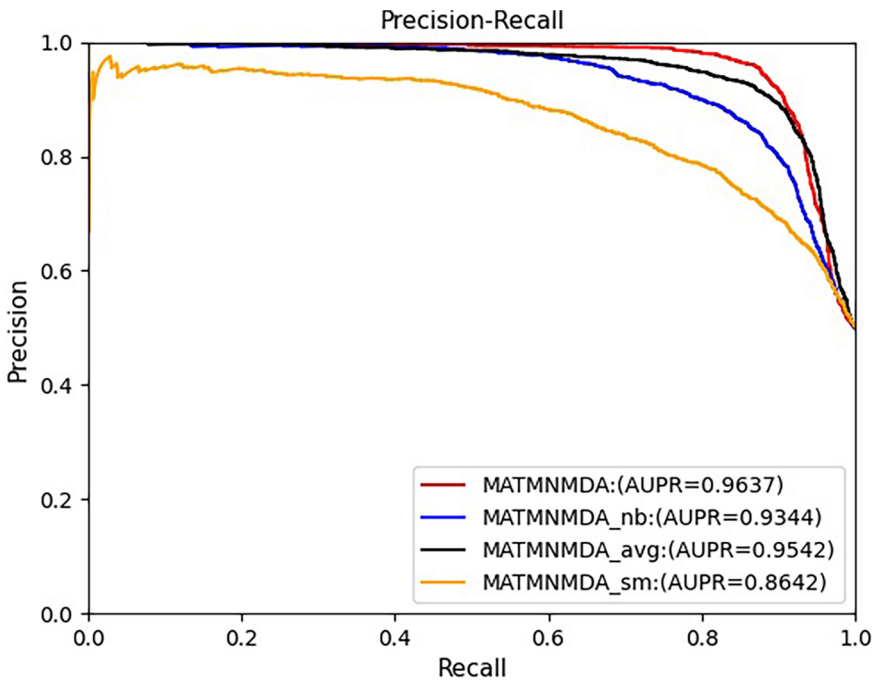

FIGURE 6 | Comparison of AUPR for MATMNMDA and its variants.

evaluation metrics, and its AUC and AUPR reach 0.9492 and 0.9637, respectively, which are better than all baseline methods. The CRPGCN model occupies the second position. It applies the RWR algorithm, which allows each calculated node to better fuse information from neighboring nodes with higher weights, so that GCN can learn features faster and get higher prediction scores. Next is the LRLSHMDA model, because the topological structure in the microbe-disease association network helps the model to effectively use the hidden information of vertices and edges, which helps to train the optimal classifier, so that microbe-disease

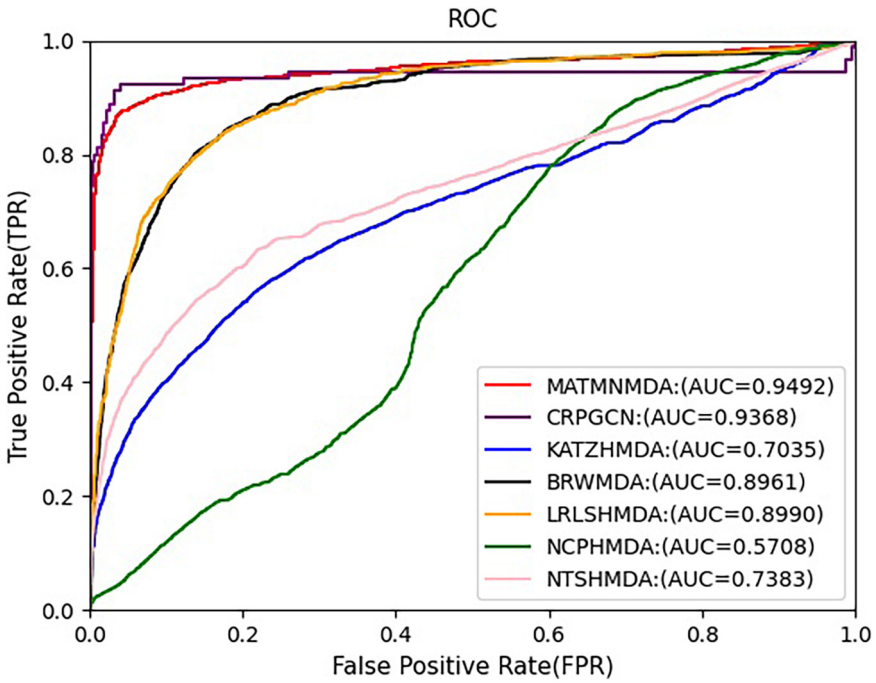

FIGURE 7 | Comparison of AUC for MATMNMDA and baselines.

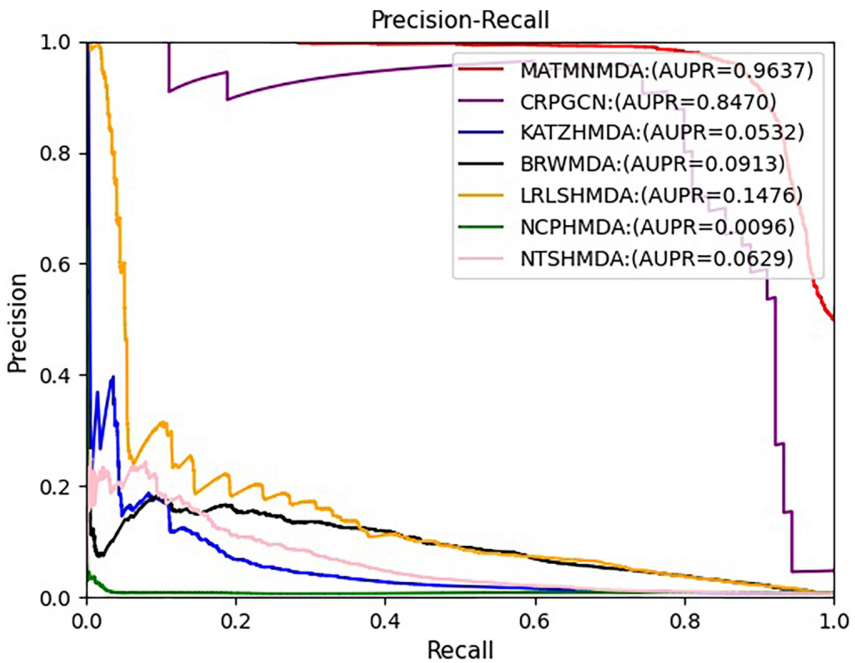

FIGURE 8 | Comparison of AUPR for MATMNMDA and baselines.

associations can be predicted more accurately. Next is the BRWMDA model, which also achieved good prediction results, because the BRWMDA model is based on similarity and bi-random walk, and it can model the topology information of the network well. However, NCPHMDA and NTSHMDA have poor prediction performance, because although we have obtained 9,202 known microbe-disease associations, they account for 0.7% of the whole microbe-disease association. The whole network is very sparse, and these two methods are based on network structure, so their performance is relatively poor.

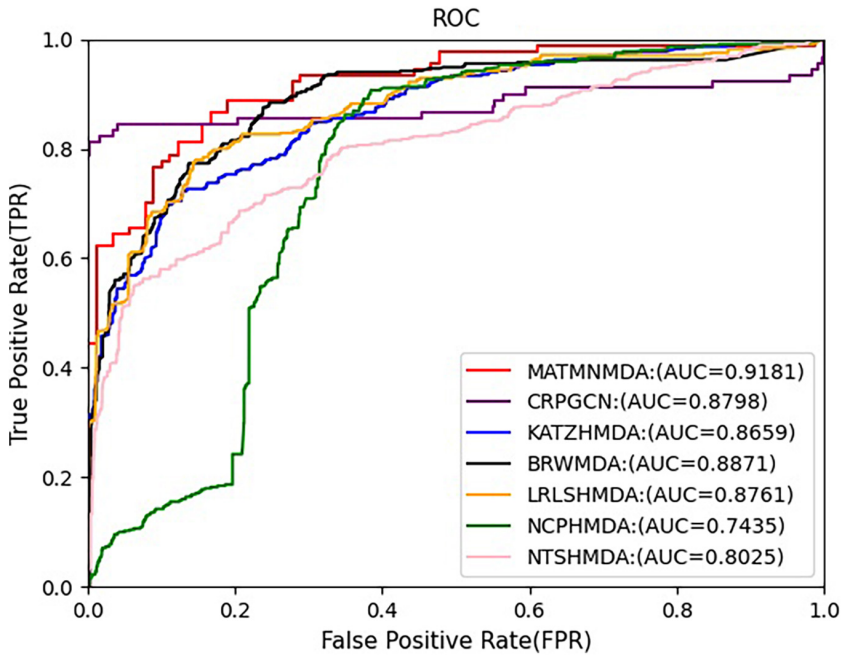

FIGURE 9 | Comparison of AUC for MATMNMDA and baselines on HMDAD dataset.

Comparison With Different Datasets

In this study, we augment the known microbe-disease association data. In order to verify the validity of the MATHNMDA model in our dataset, we also compare MATHNMDA with baseline methods on HMDAD and Disbiome, which are commonly used in microbe-disease prediction. The results are shown in **Figures 9**, **10**, and the comparison results of these methods on the three datasets are shown in **Table 1**. From **Table 1**, we can see that on each dataset, our model achieves the highest prediction value. It performs best on our dataset, so we can suggest that augmenting the known microbe-disease associations can help to improve the performance of the MATHNMDA. In addition, we can see that CRPGCN, LRLSHMDA, and BRWMDA methods perform well on these three datasets among the baseline methods. It also shows that these three methods are suitable for both large and small datasets, and the robustness of models is better. The remaining comparison methods are only suitable for small datasets.

Case Study

To further evaluate the predictive ability of the MATMNMDA model in identifying new microbe-disease associations, we conduct case studies on asthma, inflammatory bowel disease (IBD), and COVID-19. For each disease, microbes that have known associations with the disease are first removed. Then the predicted scores of candidate microbes are sorted in descending order according to the MATMNMDA model. Finally, we verify whether the top 10 microbes associated with the disease are confirmed by the relevant literature.

Asthma is a heterogeneous disease characterized by chronic airway inflammation (Lee and Kim, 2021). More than 300 million people worldwide suffer from asthma, and the incidence of asthma increased by 12.6% between 1990 and 2015 (Vasily, 2017). Therefore, it is necessary to study asthma deeply. With the development of 16rRNA sequencing technology, it has been found that there is an important relationship between asthma and microbe. In this study, when we employ the MATMNMDA model to predict potential microbe-disease associations, 7 of the top 10 candidate microbes are verified by relevant literature in PubMed (as shown in **Table 2**). For example, studies have shown that *Staphylococcus* (2nd) is linked to asthma attacks (Zhou et al., 2019), the relative abundance of *Bacteroidetes, Clostridium* (3rd), and *Enterobacteriaceae* were high, and the relative abundance of *Bifidobacterium* and *Lactobacteriaceae* were low, which is associated with allergies, eczema, or asthma (Zimmermann P. et al., 2019). An increased prevalence of *Staphylococcus aureus* (6th) colonization and sensitivity against its proteins are found in asthma (Tomassen et al., 2013). Bacterial dysbiosis and abundance within *Firmicutes* (4th) were significantly reduced in asthmatic children (Hufnagl et al., 2020). *Human parainfluenza virus* 1 (4th) was detected most frequently from patients with URI (3.74%, 47/1,257), followed by those with bronchitis (2.14%, 53/2,479), pneumonia (0.85%, 145/17,068), bronchiolitis (0.47%, 12/2,536), and asthma (0.43%, 2/462; Wang et al., 2015). *Herpesviruses* were the most abundant virus type in the asthma group (44.6 $\pm$ 4.6%), mainly *cytomegalovirus* (CMV; 9th) and EBV, which accounted for 24.5 $\pm$ 3.3 and 16.9 $\pm$ 3.5%, respectively (Choi et al., 2021). In healthy controls, the two viruses were 5.4 $\pm$ 2.5 and 7.1 $\pm$ 3.0%, respectively. Therefore, CMV and EBV are more abundant in patients with asthma exacerbations.

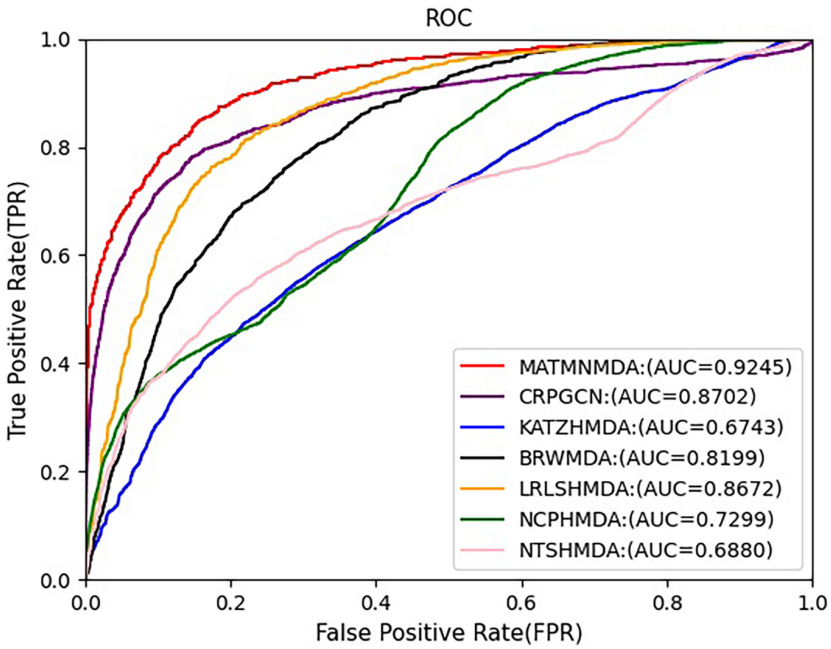

FIGURE 10 | Comparison of AUC for MATMNMDA and baselines on Disbiome dataset.

Inflammatory bowel disease (IBD) is an idiopathic intestinal inflammatory disease, mainly including ulcerative colitis (UC) and Crohn's disease (CD), with clinical manifestations of abdominal pain, diarrhea, and bloody stools. It is difficult to completely cure the disease, which is easy to recur, and there is a potential risk of cancer. Therefore, we perform a case study of IBD to evaluate the predictive ability of the MATMNMDA model for novel microbe-disease associations. The results are shown in **Table 3**, and 7 of the top 10 candidate microbes are verified by relevant literature. For example, *Fusobacterium* (2nd), *Halomonas*, *Acinetobacter*, *Shewanella*, and *Streptococcus* were enriched in the CD microbiota (Weng et al., 2019). Increased abundance of *Salmonella* sp., *Campylobacter* sp., *Helicobacter* sp., *Escherichia coli*, *Alcaligenes* sp., and *Mycobacterium* sp. (4th) was observed in IBD patients (Olejniczak-Staruch et al., 2021). IBD patients exhibit a lower abundance of butyrate-producing bacteria (6th; Gasaly et al., 2021) and butyrate content. Although some findings related to IBD dysbiosis have varied among the studies due to differences in sample type, survey method, patient profile, and drug treatment, the most consistent observation across these studies is that bacterial diversity decreased in IBD patients. For viruses infecting human cells, *Anelloviridae* (5th) showed a higher prevalence in very early-onset IBD compared to healthy controls (Liang et al., 2020). The population of *Firmicutes* decreased (7th) and that of *Proteobacteria* increased (Matsuoka and Kanai, 2015). Researchers observed a bias in the fungal microbiota in IBD compared to the normal control group, with an increased *Basidiomycota*/*Ascomycota* ratio (8th), a decreased *Saccharomyces cerevisiae* ratio, and an increased *Candida albicans* ratio (Sokol et al., 2017). There are experiments to verify that the intensity of both CMV and *human herpesvirus* 6 (HHV-6; 9th) correlated with endoscopic disease severity in IBD (CMV, $p = 0.010$ and HHV-6, $p = 0.048$; Sipponen et al., 2011).

Coronavirus disease 2019 (COVID-19) is a disease caused by severe respiratory syndrome coronavirus 2 (SARS-CoV-2). It has

TABLE 1 | Performance comparison of MATMNMDA and baselines on different datasets.

DATASET	HMDAD		Disbiome		Our dataset	
METHOD	AUC	AUPR	AUC	AUPR	AUC	AUPR
CRPGCN	0.8798	0.4533	0.8702	0.4965	0.9368	0.8470
KATZHMDA	0.8815	0.4828	0.6743	0.0508	0.7035	0.0532
BRWMDA	0.8748	0.3966	0.8199	0.0705	0.8961	0.0913
LRLSHMDA	0.8766	0.4960	0.8672	0.1370	0.8990	0.1476
NCPHMDA	0.7524	0.0795	0.7299	0.1024	0.5708	0.0092
NTSHMDA	0.8276	0.2975	0.6880	0.0630	0.7383	0.0629
MATMNMDA	0.9181	0.9297	0.9245	0.9322	0.9492	0.9637

TABLE 2 | Top 10 candidate microbes related to asthma.

Rank	Microbe	Evidence
1	Geotrichum sp.	PMID: 9376049
2	Staphylococcus	PMID: 24117882
3	Clostridiaceae	PMID: 30600099
4	Firmicutes	PMID: 32072252
5	Mitsuokella	Unconfirmed
6	Staphylococcus aureus	PMID: 30193937
7	Human parainfluenza virus 1	PMID: 26481737
8	Sphingobacteriia	Unconfirmed
9	Cytomegalovirus	PMID: 33757721
10	Aeromonas hydrophila	Unconfirmed

TABLE 3 | Top 10 candidate microbes related to IBD.

Rank	Microbe	Evidence
1	Holophagae	Unconfirmed
2	Fusobacterium	PMID: 31240835
3	Sneathia sanguinegens	Unconfirmed
4	Mycobacterium sp.	PMID: 33924414
5	Anelloviridae	PMID: 32406906
6	Butyrate-producing bacterium	PMID: 33802759
7	Firmicutes	PMID: 25420450
8	ascomycota	PMID: 26843508
9	Human herpesvirus 6	PMID: 21879802
10	Nitrososphaeraceae	Unconfirmed

TABLE 4 | Top 10 candidate microbes related to COVID-19.

Rank	Microbe	Evidence
1	Dyella	Unconfirmed
2	Acinetobacter calcoaceticus	Unconfirmed
3	Coriobacteriaceae bacterium	Unconfirmed
4	Bacteroides intestinalis	Unconfirmed
5	Bacteroides thetaiotaomicron	PMID: 32442562
6	Pisolithaceae	Unconfirmed
7	Pigmentiphaga	Unconfirmed
8	Mucor	PMID: 34009676
9	Prevotella disiens	PMID: 33577896
10	Blumeria graminis	Unconfirmed

been 3 years since its emergence and has become a pandemic threat to human health and the world economy. Although most cases of COVID-19 are mild or moderate, 3–4% of patients may be severe or critical, leading to hospitalization, respiratory failure, or death (Shen et al., 2020; Taleghani and Taghipour, 2021). Recent studies have found significant changes in the gut microbiome after infection with SARS-CoV-2. Therefore, this study conducts a case study on COVID-19 to evaluate the predictive ability of the model for COVID-19-related microbes, thereby helping researchers to conduct experimental verification purposefully, thus saving manpower and material resources. The results are presented in **Tables 3**, **4**, of the top 10 candidate microbes were verified by relevant literature. For example, the analysis of fecal samples from COVID-19 patients found that the populations of *Bacteroides dorei*, *Bacteroides thetaiotaomicron* (5th), *Bacteroides massiliensis*, and *Bacteroides ovatus* were negatively associated with SARS-CoV-2 viral load in the samples (Zuo et al., 2020). Mycological analysis revealed that 77.8 and 30.6% of patients were infected with *Mucor* (8th) and *Aspergillus*, respectively (El-Kholy and El-Fattah, 2021). *Staphylococcus haemolyticus*, *Prevotella disiens* (9th), and 2 *Corynebacterium*_1 unclassified amplicon sequence variants were more abundant in people with low SARS-CoV-2 viral load during COVID-19 infection (Rosas-Salazar et al., 2021).

CONCLUSION

Increasing studies have shown that microbes play a key role in human health and disease. Microbe-disease associations cannot only reveal disease pathogenesis, but also promote the diagnosis and prognosis of diseases. Therefore, research on microbe-disease associations has attracted wide attention. In this study, we propose a novel computational model, called MATMNMDA, to predict potential microbe-disease associations. In order to capture more semantic and structural information between microbe nodes and disease nodes, we introduce drugs to construct a tripartite heterogeneous network and apply MAGNN to learn low-dimensional embedded representations of microbe nodes and disease nodes. For each layer of MAGNN, we use intra-metapath aggregation to get the representation of the target node in each metapath, which is the input of inter-metapath aggregation layer. Then we aggregate the embedding representations between different metapaths related to the target node. Therefore, we can learn the embedding representation for the target node (microbe node or disease node) of the layer. Finally, we obtain vector representations of microbes and diseases based on the output of the last layer in the MAGNN, which is used for the prediction task. We designed multiple experiments to verify the effectiveness of the MATMNMDA model. By analyzing the experimental results, we found that: (1) Compared to the variants of our model, our model obtains the best prediction performance, which also indicates that our method could be better applied to microbe-disease prediction. (2) Under the same conditions, compared to the state-of-the-art methods, our method also obtains the best AUC and AUPR, which indicates that the MATMNMDA model can better identify potential disease-related microbes. (3) Compared to the state-of-the-art methods on different datasets, MATMNMDA achieves the best prediction performance on our enlarged dataset. It demonstrates that more known microbe-disease associations can help MATHMDA improve predictive performance. (4) Case studies on asthma, IBD, and COVID-19 further verified the effectiveness of MATMNMDA.

In future work, we will add more relational data, such as drug-drug interactions, drug-protein interactions, and protein-disease associations to achieve better predictive results.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

XL conceived, designed and managed the study, analyzed the results, and revised the manuscript. YC conducted the experiments, analyzed the results, and wrote the manuscript. Both authors read and approved the final manuscript.

FUNDING

This work was supported by the National Natural Science Foundation of China (61972451 and 61902230).

REFERENCES

Akanksha, R., Anamika, T., Shivangi, S., and Manoj, K. (2017). A biofilm: a resource of anti-biofilm agents and their potential implications in targeting antibiotic drug resistance. *Nucleic Acids Res.* 46, D894–D900. doi: 10.1093/nar/gkx 1157

Andersen, P. I., Ianevski, A., Lysvand, H., Vitkauskiene, A., and Kainov, D. E. (2020). Discovery and development of safe-in-man broad-spectrum antiviral agents. *Int. J. Infect. Dis.* 93, 268–276. doi: 10.1016/j.ijid.2020.02.018

Arrieta, M., Stiemsma, L., Dimitriu, P., Thorson, L., Russell, S., Yurist-Doutsch, S., et al. (2015). Early infancy microbial and metabolic alterations affect risk of childhood asthma. *Sci. Transl. Med.* 7:307ra152. doi: 10.1126/scitranslmed.aab2271

Bian, C., Lei, X., and Wu, F. (2021). GATCDA: Predicting circRNA-Disease Associations Based on Graph Attention Network. *Cancers* 13:2595. doi: 10.3390/cancers13112595

Chen, L., Zheng, D., Liu, B., Jian, Y., and Jin, Q. (2016). VFDB 2016: hierarchical and refined dataset for big data analysis—10 years on. *Nucleic Acids Res.* 44, D694–D697. doi: 10.1093/nar/gkv1239

Chen, X., Huang, Y., You, Z., Yan, G., and Wang, X. (2016). A novel approach based on KATZ measure to predict associations of human microbiota with non-infectious diseases. *Bioinformatics* 33, 733–739.

Choi, S., Sohn, K. H., Jung, J. W., Kang, M. G., Yang, M. S., Kim, S., et al. (2021). Lung virome: new potential biomarkers for asthma severity and exacerbation. *J. Allergy Clin. Immun.* 148, 1007–1015. doi: 10.1016/j.jaci.2021.03.017

Davis, A. P., Murphy, C. G., Johnson, R., Lay, J. M., Lennon-Hopkins, K., Saraceni-Richards, C., et al. (2012). The Comparative Toxicogenomics Database: update 2013. *Nucleic Acids Res.* 41, D1104–D1114. doi: 10.1093/nar/gks994

El-Kholy, N. A., and El-Fattah, A. M. A. (2021). Invasive Fungal Sinusitis in Post COVID-19 Patients: A New Clinical Entity. *Laryngoscope* 131, 2652–2658. doi: 10.1002/lary.29632

Fan, C., Lei, X., and Guo, L. (2018). Predicting the associations between microbes and diseases by integrating multiple data sources and path-based HeteSim scores. *Neurocomputing* 323, 76–85.

Gasaly, N., Hermoso, M. A., and Gotteland, M. (2021). Butyrate and the Fine-Tuning of Colonic Homeostasis: Implication for Inflammatory Bowel Diseases. *Int. J. Mol. Sci*. 22:3061. doi: 10.3390/ijms22063061

Gill, S., Pop, M., DeBoy, R., Eckburg, P., Turnbaugh, P., Samuel, B., et al. (2006). Metagenomic analysis of the human distal gut microbiome. *Science* 312, 1355–1359. doi: 10.1126/science.1124234

He, B., Peng, L., and Li, Z. (2018). Human Microbe-Disease Association Prediction With Graph Regularized Non-Negative Matrix Factorization. *Front. Microbiol.* 9:2560. doi: 10.3389/fmicb.2018.02560

Huang, Z., Chen, X., Zhu, Z., Liu, H., and Wen, Z. (2017). PBHMDA: Path-Based Human Microbe-Disease Association prediction. *Front. Microbiol.* 8:233. doi: 10.3389/fmicb.2017.00233

Hufnagl, K., Pali-Schöll, I., Roth-Walter, F., and Jensen-Jarolim, E. (2020). Dysbiosis of the gut and lung microbiome has a role in asthma. *Semin. Immunopathol.* 42, 75–93. doi: 10.1007/s00281-019-00775-y

Human Microbiome Project (HMP), C. (2012). Structure, function and diversity of the healthy human microbiome. *Nature* 486, 207–214. doi: 10.1038/nature11234

Kong, W., Cui, Q., Zhou, X., Ma, Z., and Lu, Y. (2017). An analysis of human microbe-disease associations. *Brief. Bioinformatics.* 18, 85–97.

Lee, Y., and Kim, J. (2021). Urine Microbe-Derived Extracellular Vesicles in Children With Asthma. *Allergy Asthma Immunol. Res.*13, 75–87. doi: 10.4168/aair.2021.13.1.75

Lei, X., Bian, C., and Pan, Y. (2021). Predicting CircRNA-disease associations based on improved weighted biased meta-structure. *J. Comput. Sci. Technol.* 36, 288–298.

Lei, X., and Wang, Y. (2020). Predicting Microbe-Disease Association by Learning Graph Representations and Rule-Based Inference on the Heterogeneous Network. *Front. Microbiol.* 11:579. doi: 10.3389/fmicb.2020.00579

Liang, G., Conrad, M., Kelsen, J., Kessler, L., Breton, J., Albenberg, L., et al. (2020). Dynamics of the Stool Virome in Very Early-Onset Inflammatory Bowel Disease. *J. Crohns Colitis* 14, 1600–1610. doi: 10.1093/ecco-jcc/jjaa094

Liu, W., Jiang, Y., Peng, L., Sun, X., Gan, W., Zhao, Q., et al. (2022a). Inferring gene regulatory networks using the improved Markov blanket discovery algorithm. *Interdiscip. Sci.* 14, 168–181. doi: 10.1007/s12539-021-00478-9

Liu, W., Lin, H., Huang, L., Peng, L., Tang, T., Zhao, Q., et al. (2022b). Identification of miRNA–disease associations via deep forest ensemble learning based on autoencoder. *Brief. Bioinform.* [Epub ahead of print]. doi: 10.1093/bib/bbac104

Luo, J., and Long, Y. (2018). NTSHMDA: Prediction of Human Microbe-Disease Association based on Random Walk by Integrating Network Topological Similarity. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 17, 1341–1351. doi: 10.1109/TCBB.2018.2883041

Ma, Z., Kuang, Z., and Deng, L. (2021). CRPGCN: predicting circRNA-disease associations using graph convolutional network based on heterogeneous network. *BMC Bioinform.* 22:551. doi: 10.1186/s12859-021-04467-z

Marco, M. L., Heeney, D., Binda, S., Cifelli, C. J., Cotter, P. D., Foligne, B., et al. (2017). Health benefits of fermented foods: microbiota and beyond. *Curr. Opin. Biotech.* 44, 94–102. doi: 10.1016/j.copbio.2016.11.010

Matsuoka, K., and Kanai, T. (2015). The gut microbiota and inflammatory bowel disease. *Semin. Immunopathol.* 37, 47–55.

Olejniczak-Staruch, I., Ciążyńska, M., Sobolewska-Sztychny, D., Narbutt, J., Skibińska, M., and Lesiak, A. (2021). Alterations of the Skin and Gut Microbiome in Psoriasis and Psoriatic Arthritis. *Int. J. Mol. Sci.* 22:3998. doi: 10.3390/ijms22083998

Pan, Y., Lei, X., and Zhang, Y. (2022). Association predictions of genomics, proteinomics, transcriptomics, microbiome, metabolomics, pathomics, radiomics, drug, symptoms, environment factor, and disease networks: a comprehensive approach. *Med. Res. Rev.* 42, 441–461. doi: 10.1002/med.21847

Phaisangittisagul, E. (2016). "An Analysis of the Regularization Between L2 and Dropout in Single Hidden Layer Neural Network," in *International Conference on Intelligent Systems* (Bangkok: IEEE).

Rosas-Salazar, C., Kimura, K. S., Shilts, M. H., Strickland, B. A., Freeman, M. H., Wessinger, B. C., et al. (2021). SARS-CoV-2 infection and viral load are associated with the upper respiratory tract microbiome. *J. Allergy Clin. Immunol.* 147, 1226–1233.e2. doi: 10.1016/j.jaci.2021.02.001

Sender, R., Fuchs, S., and Milo, R. (2016). Revised Estimates for the Number of Human and Bacteria Cells in the Body. *PLoS Biol.* 14:e1002533. doi: 10.1371/journal.pbio.1002533

Shen, Y., Zheng, F., Sun, D., Ling, Y., and Chen, J. (2020). Epidemiology and clinical course of COVID-19 in Shanghai, China. *Emerg. Microbes Infect.* 9, 1537–1545. doi: 10.1080/22221751.2020.1787103

Shen, Z., Jiang, Z., and Bao, W. (2017). "CMFHMDA: Collaborative Matrix Factorization for Human Microbe-Disease Association Prediction," in *International Conference on Intelligent Computing* (Cham: Springer). doi: 10.3389/fmicb.2022.834982

Sipponen, T., Turunen, U., Lautenschlager, I., Nieminen, U., Arola, J., and Halme, L. (2011). Human herpesvirus 6 and cytomegalovirus in ileocolonic mucosa in inflammatory bowel disease. *Scand. J. Gastroenterol.* 46, 1324–1333. doi: 10.3109/00365521.2011.605466

Skoufos, G., Kardaras, F., Alexiou, A., Kavakiotis, I., and Hatzigeorgiou, A. (2020). Peryton: a manual collection of experimentally supported microbe-disease associations. *Nucleic Acids Res.* 49, D1328–D1333. doi: 10.1093/nar/gkaa902

Sokol, H., Leducq, V., Aschard, H., Pham, H., Jegou, S., Landman, C., et al. (2017). Fungal microbiota dysbiosis in IBD. *Gut* 66, 1039–1048. doi: 10.1136/gutjnl-2015-310746

Sun, Y., Zhang, D., Cai, S., Ming, Z., and Li, J. (2018). MDAD: A Special Resource for Microbe-Drug Associations. *Front. Cell. Infect. Microbiol.* 8:424. doi: 10.3389/fcimb.2018.00424

Sun, Z., Deng, Z., Nie, J., and Tang, J. (2019). "*RotatE: Knowledge Graph Embedding by Relational Rotation in Complex Space," in *7th International Conference on Learning Representations* (New Orleans, Louisiana, United States: ICLR 2019).

Taleghani, N., and Taghipour, F. (2021). Diagnosis of COVID-19 for controlling the pandemic: a review of the state-of-the-art. *Biosens. Bioelectron.* 174:112830. doi: 10.1016/j.bios.2020.112830

Tomassen, P., Jarvis, D., Newson, R., Van Ree, R., Forsberg, B., Howarth, P., et al. (2013). Staphylococcus aureus enterotoxin-specific IgE is associated with asthma in the general population: a GA(2)LEN study. *Allergy* 68, 1289–1297. doi: 10.1111/all.12230

Vasily, V. (2017). Global, regional, and national deaths, prevalence, disability-adjusted life years, and years lived with disability for chronic obstructive pulmonary disease and asthma, 1990-2015: a systematic analysis for the Global Burden of Disease Study 2015. *Lancet Respir. Med.* 5, 691–706. doi: 10.1016/S2213-2600(17)30293-X

Wang, C., Han, C., Zhao, Q., and Chen, X. (2021). Circular RNAs and complex diseases: from experimental results to computational models. *Brief. Bioinform.* 22:bbab286. doi: 10.1093/bib/bbab286

Wang, F., Huang, Z., Chen, X., Zhu, Z., Wen, Z., and Zhao, J. (2017). LRLSHMDA: Laplacian Regularized Least Squares for Human Microbe–Disease Association prediction. *Sci. Rep.* 7:7601. doi: 10.1038/s41598-017-08127-2

Wang, F., Zhao, L., Zhu, R., Deng, J., Sun, Y., Ding, Y., et al. (2015). Parainfluenza Virus Types 1, 2, and 3 in Pediatric Patients with Acute Respiratory Infections in Beijing During 2004 to 2012. *Chin. Med. J.* 128, 2726–2730. doi: 10.4103/0366-6999.167297

Wang, L., Tan, Y., Yang, X., Kuang, L., and Ping, P. (2022). Review on predicting pairwise relationships between human microbes, drugs and diseases: from biological data to computational models. *Brief. Bioinform.* [Epub ahead of print]. doi: 10.1093/bib/bbac080

Wang, T., Cai, G., Qiu, Y., Fei, N., Zhang, M., Pang, X., et al. (2012). Structural segregation of gut microbiota between colorectal cancer patients and healthy volunteers. *ISME J.* 6, 320–329. doi: 10.1038/ismej.2011.109

Wang, Y., Lei, X., and Pan, Y. (2022). Predicting Microbe-disease Association Based on Heterogeneous Network and Global Graph Feature Learning. *Chin. J. Electron.* 31, 1–9.

Wen, Z., Yan, C., Duan, G., Li, S., Wu, F., and Wang, J. (2021). A survey on predicting microbe-disease associations: biological data and computational methods. *Brief. Bioinform.* 22:bbaa157. doi: 10.1093/bib/bbaa157

Weng, Y., Gan, H., Li, X., Huang, Y., Li, Z., Deng, H., et al. (2019). Correlation of diet, microbiota and metabolite networks in inflammatory bowel disease. *J. Dig. Dis.* 20, 447–459. doi: 10.1111/1751-2980.12795

Wishart, D. S., Feunang, Y. D., Guo, C. A., Lo, E. J., and Wilson, M. (2017). DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res.* 46, D1074–D1082. doi: 10.1093/nar/gkx1037

Wu, C., Gao, R., Zhang, D., Han, S., and Zhang, Y. (2018). PRWHMDA: Human Microbe-Disease Association Prediction by Random Walk on the Heterogeneous Network with PSO. *Int. J. Biol. Sci.* 14, 849–857. doi: 10.7150/ijbs.24539

Yan, C., Duan, G., Wu, F., Pan, Y., and Wang, J. (2019). BRWMDA:Predicting microbe-disease associations based on similarities and bi-random walk on disease and microbe networks. *IEEE ACM Trans. Comput. Biol. Bioinform*. 17, 1595–1604. doi: 10.1109/TCBB.2019.2907626

Yang, L., Li, L., and Yi, H. (2022). DeepWalk based method to predict lncRNA-miRNA associations via lncRNA-miRNA-disease-protein-drug graph. *BMC Bioinform.* 22:621. doi: 10.1186/s12859-022-04579-0

Yao, G., Zhang, W., Yang, M., Yang, H., and Li, W. (2021). MicroPhenoDB Associates Metagenomic Data with Pathogenic Microbes, Microbial Core Genes, and Human Disease Phenotypes. *Genom. Proteom. Bioinform*. 18, 760–772. doi: 10.1016/j.gpb.2020.11.001

Yorick, J., Joachim, N., Antoon, B., Nathan, D., Frederick, V., Evelien, W., et al. (2018). Disbiome database: linking the microbiome to disease. *BMC Microbiol.* 18:50. doi: 10.1186/s12866-018-1197-5

Zhang, L., Yang, P., Feng, H., Zhao, Q., and Liu, H. (2021). Using Network Distance Analysis to Predict lncRNA–miRNA Interactions. *Interdiscip. Sci.* 13, 535–545. doi: 10.1007/s12539-021-00458-z

Zhang, Y., Lei, X., Fang, Z., and Pan, Y. (2020). CircRNA-disease associations prediction based on metapath2vec++ and matrix factorization. *Big Data Mining Anal.* 3, 280–291.

Zhou, Y., Jackson, D., Bacharier, L. B., Mauger, D., Boushey, H., Castro, M., et al. (2019). The upper-airway microbiota and loss of asthma control among asthmatic children. *Nat. Commun.* 10:5714. doi: 10.1038/s41467-019-13698-x

Zhu, S., Bing, J., Min, X., Lin, C., and Zeng, X. (2018). Prediction of Drug-Gene Interaction by Using Metapath2vec. *Front. Genet.* 9:248. doi: 10.3389/fgene.2018.00248

Zimmermann, M., Patil, K. R., Typas, A., and Maier, L. (2021). Towards a mechanistic understanding of reciprocal drug-microbiome interactions. *Mol. Syst. Biol.* 17:e10116. doi: 10.15252/msb.202010116

Zimmermann, M., Zimmermann-Kogadeeva, M., Wegmann, R., and Goodman, A. L. (2019). Mapping human microbiome drug metabolism by gut bacteria and their genes. *Nature* 570:1. doi: 10.1038/s41586-019-1291-3

Zimmermann, P., Messina, N., Mohn, W. W., Finlay, B. B., and Curtis, N. (2019). Association between the intestinal microbiota and allergic sensitization, eczema, and asthma: a systematic review. *J. Allergy Clin. Immunol.* 143, 467–485. doi: 10.1016/j.jaci.2018.09.025

Zou, S., Zhang, J., and Zhang, Z. (2018). Novel human microbe-disease associations inference based on network consistency projection. *Sci. Rep.* 8:8034. doi: 10.1038/s41598-018-26448-8

Zuo, T., Zhang, F., Lui, G. C. Y., Yeoh, Y. K., Li, A. Y. L., Zhan, H., et al. (2020). Alterations in Gut Microbiota of Patients With COVID-19 During Time of Hospitalization. *Gastroenterology* 159, 944–955.e8. doi: 10.1053/j.gastro.2020.05.048

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.