




Microbe-Disease Association Prediction Using RGCN Through Microbe-Drug-Disease Network

Yueyue Wang , Xiujuan Lei , and Yi Pan 

Abstract—Accumulating evidence has shown that microbes play significant roles in human health and diseases. Therefore, identifying microbe-disease associations is conducive to disease prevention. In this article, a predictive method called TNRGCN is designed for microbe-disease associations based on Microbe-Drug-Disease Network and Relation Graph Convolutional Network (RGCN). First, considering that indirect links between microbes and diseases will be increased by introducing drug related associations, we construct a Microbe-Drug-Disease tripartite network through data processing from four databases including Human Microbe-Disease Association Database (HMDAD), Disbiome Database, Microbe-Drug Association Database (MDAD) and Comparative Toxicogenomics Database (CTD). Second, we construct similarity networks for microbes, diseases and drugs via microbe function similarity, disease semantic similarity and Gaussian interaction profile kernel similarity, respectively. Based on the similarity networks, Principal Component Analysis (PCA) is utilized to extract main features of nodes. These features will be input into the RGCN as initial features. Finally, based on the tripartite network and initial features, we design two-layer RGCN to predict microbe-disease associations. Experimental results indicate that TNRGCN achieves best performance in cross validation compared with other methods. Meanwhile, case studies for Type 2 diabetes (T2D), Bipolar disorder and Autism demonstrate the favorable effectiveness of TNRGCN in association prediction.

Index Terms—Autism, bipolar disorder, microbe-disease associations, microbe-drug-disease network, relation graph convolutional network, type 2 diabetes.

I. INTRODUCTION

Microbe communities are tiny organisms mainly including eukaryotes, archaea, bacteria and viruses, which are regarded as a special “organ” of human beings [1]. Thousands of microbes reside in different parts of human organs and tissues such as skin, oral cavity and gastrointestinal tract. In general, microbes are harmless to human health, and even have a beneficial side. For example, a variety of oral microbiomes residing in the human oral cavity interact with each other, protecting the human body against harmful external stimulation [2]. Researchers found

that intestinal flora affects brain development such as cognitive function and basic behavior patterns, and its disorder will have a negative impact on psychological health [3].

Most microbes residing in the human body maintain homeostasis, providing ecosystem services [4]. But these balances can be easily changed by the external environment and their own environment such as diet, genotype and colonization history [5]. For example, diet regulates and supports the intestinal microbiota. Different types, qualities and sources of food affect the composition and function of intestinal microbes as well as host-microbiome interactions [6]. These balances are closely related to human diseases. For instance, in patients with inflammatory bowel disease (IBD), microbial diversity is generally reduced. On average, IBD patients carry 75% of the microbial genes of healthy people [7]. Another study found that the composition of intestinal microbial group is related to the symbol of atherosclerosis and arterial hardness [8]. Thus, knowing microbe-disease associations is beneficial to disease diagnosis and treatment.

Recently, two databases for relationships with microbes and diseases have been established: Human Microbe-Disease Association Database (HMDAD) and Disbiome. HMDAD contains 483 microbe-disease associations collected manually from 61 literatures [9]. Disbiome is a constantly updated database, including confirmed microbe-disease associations and experimental verification records from various literatures and database [10].

Based on confirmed microbe-disease associations from these databases, many predictive methods have been designed to explore more unknown associations. In 2017, Chen et al. first integrated confirmed associations and Gaussian interaction profile (GIP) similarities, to construct a microbe-disease heterogeneous network. Then, they predicted microbe-disease associations by integrating step length and the number of paths between two nodes in the heterogeneous network [11]. Inspired by this method, Li et al. reconstructed a microbe-disease heterogeneous network by integrating confirmed associations and normalized GIP kernel similarity. Then they combining bidirectional recommendations and KATZ method on the heterogeneous network [12]. Jiang et al. constructed a knowledge graph centered on microbes and diseases by collecting data from multiple databases. And then, they predicted potential associations by utilizing graph neural network method to learn nodes’ representation of knowledge graph [13]. Peng et al. predicted potential microbe-disease associations by using Kronecker sum operations and eigenvalue transformation on similarity networks [14]. Hua et al.

Manuscript received 26 February 2022; revised 11 October 2022; accepted 16 February 2023. Date of publication 22 February 2023; date of current version 26 December 2023. We thank the financial support from National Natural Science Foundation of China under Grants 62272288, 61972451, 61902230 and U22A2041. (Corresponding author: Xiujuan Lei.)

Yueyue Wang and Xiujuan Lei are with the School of Computer Science, Shaanxi Normal University, Xi’an, Shaanxi 710119, China (e-mail: yueyue-wang@snnu.edu.cn; xjlei@snnu.edu.cn).

Yi Pan is with the Faculty of Computer Science and Control Engineering, Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen, Guangdong 518055, China (e-mail: yipan@gsu.edu).

Digital Object Identifier 10.1109/TCBB.2023.3247035

utilized Graph Augmentation Convolutional Network and attention mechanism to learn node features. And then they combined matrix completion to predict potential associations [15]. Li et al. constructed a three-layer back-propagation neural network with a new activation function based on hyperbolic tangent function. In order to improve operating efficiency, they used microbe GIP kernel similarity to weight the initial connection value [16]. Wang et al. designed a predicted method called MSLINE. This method learned the multi-level domain information in microbe-disease network by combining an embedding algorithm LINE and random walk [17]. Liu et al. explored microbe-disease associations by utilizing a multi-component graph attention network and a fully connected network [18]. Peng et al. aggregated multi-view features including linear feature and nonlinear feature. This method takes into account the complementarity of different features [19]. Chen et al. constructed a microbe-drug-disease heterogeneous network, and then used the multi-head attention mechanism to aggregate different meta-paths to learn the nodes' features [20].

These methods achieved good predictive performance, but currently used database contain few microbe-disease associations. For example, HMDAD contains only 39 diseases, and using this database alone will result in too few diseases that can be predicted. Moreover, many different databases including microbes, diseases and drugs have been constructed, such as Microbe-Drug Association Database (MDAD) and Comparative Toxicogenomics Database (CTD). these databases contain various types of nodes, which can directly or indirectly increase the relationship between microbes and diseases, and are conducive to prediction.

Graph Convolution Network (GCN) aggregates neighbors' information through convolutional operation to extracts node features in a network [21]. It has been widely used and has shown good performance in association prediction. In the field of bioinformatics, GCN has been applied to predict circRNA-disease association [22], metabolite-disease association [23] and drug-drug association [24]. However, in the process of aggregation, it treats all neighbor nodes as the same type, and does not selectively aggregate according to the type of neighbor nodes. Considering this reason, Relation Graph Convolutional Network (RGCN) [25] considers neighbor node types and the connection direction with the current node when convolution. Therefore, it can be applied to heterogeneous networks with different types of nodes and edges.

In this article, we predict microbe-disease association based on Microbe-Drug-Disease Tripartite Network and RGCN (called TNRGCN). First, we construct a microbe-drug-disease tripartite network by screening microbe-disease associations, microbe-drug associations and disease-drug associations from HMDAD, Disbiome, MDAD and CTD. Second, we use Principal Component Analysis (PCA) to extract main features of nodes in similarity networks, and input them into RGCN as initial features. These similarity networks include microbe function similarity, disease semantic similarity and GIP similarity. Finally, based on the microbe-drug-disease tripartite network and initial features, we utilize two-layer RGCN to predict potential associations. Compared with other methods, TNRGCN achieves

TABLE I
DETAILS OF THE FILTERED DATA

Database	Microbe	Disease	Drug	Associations
HMDAD+ Disbiome	1519	254	--	7258
MDAD	1519	--	1181	3783
CTD	--	254	1181	4552

best performance. Case studies for Type 2 diabetes (T2D), Bipolar disorder and Autism also demonstrate the good performance of TNRGCN. The flowchart of TNRGCN is shown in Fig. 1.

Our main contribution is as follows. First, we integrate HMDAD and Disbiome, including more microbe-disease associations. Second, we combine different associations among microbes, diseases and drugs, which can enrich link information in the network. Third, RGCN is utilized to learn node features in microbe-drug-disease tripartite network, considering different types of nodes and edges.

II. MATERIAL AND METHODS

A. Material

The data for this article is from four databases. The microbe-disease associations are collected from HMDAD and Disbiome. HMDAD contains 483 confirmed microbe-disease associations between 292 microbes and 39 diseases. By removing duplicate associations, we obtained 450 records among them. Meanwhile, we obtain all records from Disbiome as of December, 2020, including 1585 microbes, 353 diseases and 8695 microbe-disease associations between them. After removing the duplicate associations, we finally obtain 1416 microbes, 243 diseases and their corresponding 7052 association records. The microbe-drug associations are collected from MDAD, which includes 180 microbes, 1388 drugs and their corresponding 5055 associations. The disease-drug associations are collected from CTD, including 7119363 association relation records among 12791 drugs and 7098 diseases.

B. Methods

1) *Data Processing*: First, we integrate all microbe-disease associations in HMDAD and Disbiome and remove the duplicate records. Considering that Disbiome contains 17 types disease records, including Disease or Syndrome, Organism Function, Individual Behavior and so on. According to the UMLS CUI in DisGeNET [26], we compared disease ID with UMLS CUI, and screened out three types of diseases: Disease or Syndrome, Mental or Behavioral Dysfunction and Neoplastic Process. Finally, we obtain 254 diseases and 7258 microbe-disease associations related to them, involving 1519 microbes. Second, for 1519 microbes associated with diseases, we screen out 3783 microbe-drug associations from MDAD related to them, involving 1181 drugs. Third, for 1181 drugs and 254 diseases, we obtain 4552 disease-drug associations from CTD. The flowchart of data processing is as Fig. 2. The specific details of the filtered data are shown in Table I.

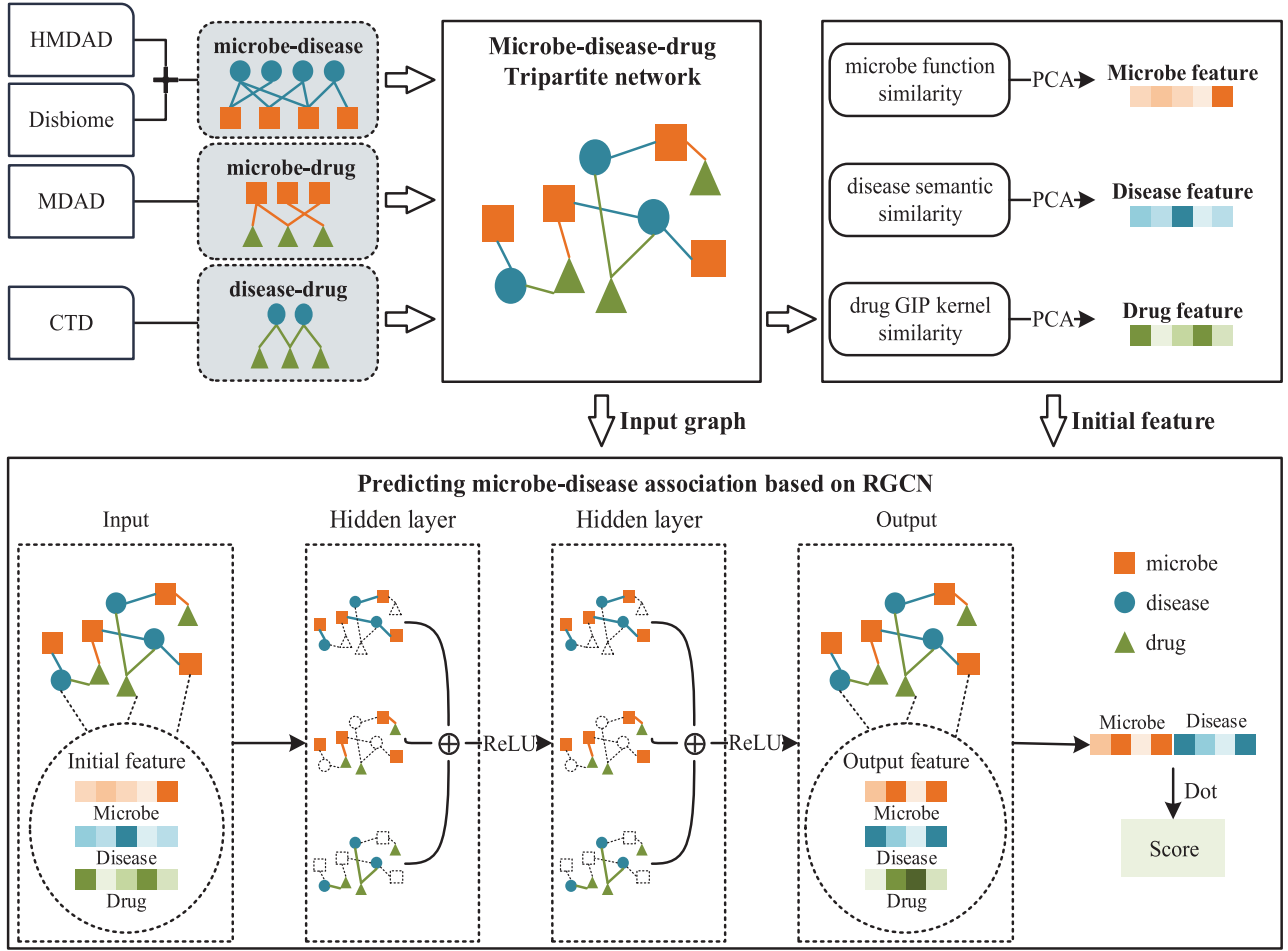


Fig. 1. Flowchart of TNRGCN.

2) *Tripartite Network Construction*: After data processing, we construct three adjacency matrices representing microbe-disease association, microbe-drug association and disease-drug association, respectively. A_{dm} represents microbe-disease associations. If disease i has association with microbe j , we set $A_{dm}(i, j) = 1$, otherwise, $A_{dm}(i, j) = 0$. Similarly, A_{mu} represents microbe-drug associations. If microbe i has association with drug j , $A_{mu}(i, j)$ is set to 1, otherwise, $A_{mu}(i, j)$ is set to 0. A_{du} represents disease-drug associations. If disease i has association with drug j , $A_{du}(i, j) = 1$, otherwise, $A_{du}(i, j) = 0$.

Since drugs are related to both microbes and diseases, we build a microbe-drug-disease tripartite network P to indirectly increase microbe-disease associations by introducing drugs.

3) *Feature Initialization*: During the learning process, we first initialize the features of the nodes. By calculating the similarity of microbes, diseases and drugs separately, and using PCA to obtain the main features of the similarity as the initial features.

1) Microbe similarity calculating

HMDAD and Disbiome contain the organs where microbes live and the effects of microbes on them. In a previous article, authors calculated microbial function similarity based on their host organs in the human body [27]. However, they only considered the location of colonization, and did not consider the regulatory

role of different microbes in the same organ. On this basis, we calculate microbe function similarity based on the assumption that microbes share stronger function similarities if they have the same effects on the same organ. Specifically, if microbe i and j live in a same organ and have same regulation (increase or decrease), we add 1 to $M_F(i, j)$. After calculating the influence of all microbes on the resident organs, we normalize microbe function similarity according to (1):

$$M_F(i, j) = \frac{M_F(i, j) - \min(M_F)}{\max(M_F) - \min(M_F)} \quad (1)$$

where $\max(M_F)$ and $\min(M_F)$ are the maximum and minimum of matrix M_F , respectively.

2) Disease similarity calculating

The disease semantic similarity is calculated according to Mesh database [28]. In Mesh, each disease is represented as a Directed Acyclic Graph (DAG), including a disease and its dependencies among all its ancestors. The contribution of every element in a DAG according to (2) [29]:

$$D_{con}(d) = \begin{cases} 1 & \text{if } d = D \\ \max \{ \Delta \times D_{con}(d') | d' \in \text{children of } d \} & \text{if } d \neq D \end{cases} \quad (2)$$

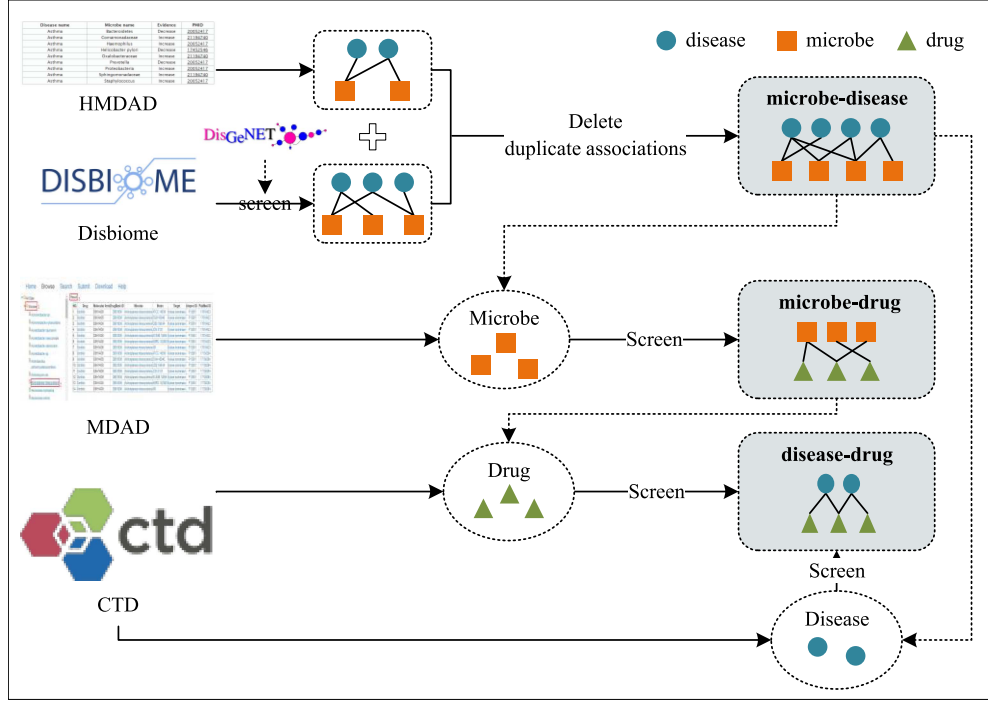


Fig. 2. Flowchart of data processing.

where Δ is a semantic contribution decay factor. The previous article usually set Δ to 0.5 [29].

The contribution to a certain disease is the sum of all the contributions of all elements in its DAG:

$$D_{tc}(d) = \sum_{t \in V_d} D_{con}(t) \quad (3)$$

where V_d contains disease d and all its ancestors.

For disease i and j , the semantic similarity of them is calculated by the contribution of them. The formula is shown as (4):

$$D_s(d_i, d_j) = \frac{\sum_{t \in V(d_i) \cap V(d_j)} D(i)_{con}(t) + D(j)_{con}(t)}{D_{tc}(i) + D_{tc}(j)} \quad (4)$$

3) Drug similarity calculating

We assume that when two drugs have more of the same neighbor nodes, their functions are more similar. Thus, we calculate the drug GIP kernel similarity G_{u1} according to the disease-drug association network. The GIP similarity of drug i and j are calculated as (5) and (6) [30]:

$$G_{u1}(i, j) = \exp(-\gamma_{u1} \|A_{du}(u(i)) - A_{du}(u(j))\|^2) \quad (5)$$

$$\gamma_{u1} = \gamma'_{u1} / \frac{1}{N_u} \sum_{i=1}^{N_u} \|A_{du}(u(i))\|^2 \quad (6)$$

where $A_{du}(u(i))$ is the i th column of matrix A_{du} . γ_{u1} is a normalized kernel bandwidth parameter affected by parameter γ'_{u1} . According to previous study, we set γ'_{u1} to 1 [30]. $N_u = 1181$, is the total number of drugs.

Similarly, the drug GIP kernel similarity G_{u2} is calculated based on microbe-drug association network. By combining the

two similarities, we obtain the drug similarity network G_u , which is shown as (7):

$$G_u = (G_{u1} + G_{u2})/2 \quad (7)$$

PCA uses linear transformation to achieve dimensionality reduction. The main idea of PCA is to map n -dimensional features to k -dimensions. By retaining the first k coordinate axes containing most of the variance, and ignoring the feature dimensions containing almost zero variance, the dimensionality reduction processing of data features is realized. Due to different dimensions of similarity between microbes, diseases and drugs, we utilize PCA to reduce the dimensionality of each node and feed them to RGCN as initial features. In this article, we set the dimension of initial features to 128.

4) Predicting Microbe-Disease Associations Based on RGCN: GCN learns features by aggregating neighbors' information through weighted summation of nodes in each layer. The hidden layer l -th representation of each node i in GCN is calculated as (8) [21]:

$$H^{(l+1)} = \sigma(\tilde{L}_{sym} W^{(l)} H^{(l)}) \quad (8)$$

in (8), $H^{(l)}$ is the hidden features of nodes l -th layer. $W^{(l)}$ is the weight influence factor of nodes in the l -th layer. In order to prevent overfitting, the parameters in the method learning process are limited by regularization terms. A commonly used regularization term is the symmetric Laplace matrix \tilde{L}_{sym} , which is shown as (9),

$$\tilde{L}_{sym} = \tilde{D}^{-\frac{1}{2}} \tilde{A} \tilde{D}^{-\frac{1}{2}} \quad (9)$$

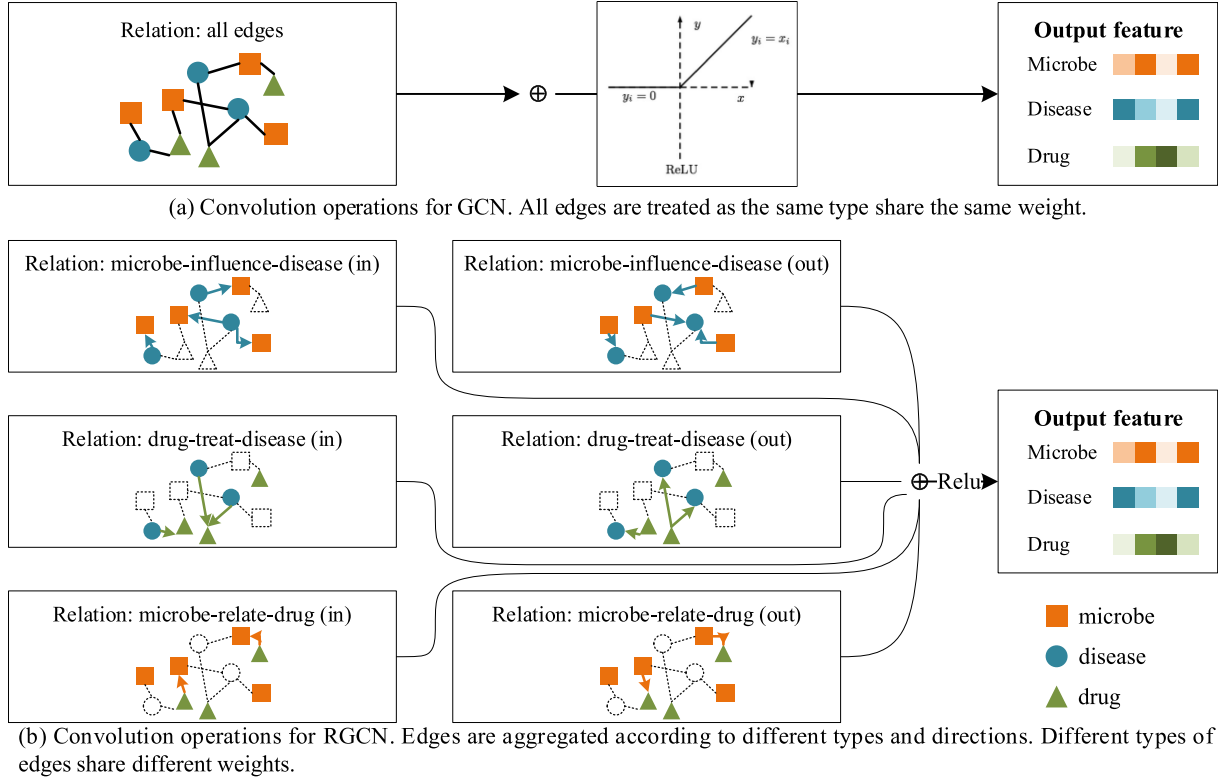


Fig. 3. Different convolution operations for GCN and RGCN in microbe-drug-disease tripartite network.

where $\tilde{A} = A + I_N$, contains the adjacency information matrix of the node and its own connection information. \tilde{D} is the degree matrix of \tilde{A} . σ is an activation function that passes information from one layer to the next layer.

Since the loop of the node itself is not included in the microbe-drug-disease tripartite network, the formula of GCN applied to the network is shown in (10):

$$H^{(l+1)} = \sigma(D^{-\frac{1}{2}} A D^{-\frac{1}{2}} W^{(l)} H^{(l)}) \quad (10)$$

where D is the degree matrix of A .

Specifically, convolution process in l -th layer is as follows:

$$h_i^{(l+1)} = \sigma \left(\sum_{j \in N_i} C_{ij} W^{(l)} h_j^{(l)} \right) \quad (11)$$

where N_i includes all neighbors of node i , $h_j^{(l)}$ is the hidden features of node j in l -th layer. $C_{ij} = N_i^{-\frac{1}{2}} N_j^{-\frac{1}{2}}$, is the product of the square root of the node degree.

GCN regards all neighbor nodes as the same type when aggregating node information. All neighbor nodes in layer l share one weight $W^{(l)}$. On the contrary, RGCN considers different types of nodes and the connection direction with the current node. different nodes and connection directions share different weights, and only edges of the same association type can use the same weight. Specifically, for a node i , the convolutional

operation of it is calculated as (12) [25]:

$$h_i^{(l+1)} = \sigma \left(\sum_{r \in R} \sum_{j \in N_i^r} C_{i,r} W_r^{(l)} h_j^{(l)} + W_0^{(l)} h_i^{(l)} \right) \quad (12)$$

where $h_i^{(l)}$ is the hidden features of node i in the l -th layer. $r \in R$ represents the type of edges. N_i^r includes all neighbors of node i under relation r . $C_{i,r} = 1/(N_i^r)$, is a regularization term. $W_r^{(l)}$ is the weight corresponding to the relation r in l -th layer. σ is an activation function. In this article, we use ReLU as the activation function.

RGCN considers both edge types and edge orientations, so in microbe-drug-disease tripartite network, we consider 6 types of edges including two directions: “microbe-influence-disease (in/out)”, “microbe-relate-drug (in/out)” and “drug-treat-disease (in/out)”. In the process of aggregation, because the microbe-drug-disease tripartite network has no self-loop edges, the convolutional operation only accumulates all the features from the neighbor nodes. The different convolution operations for GCN and RGCN in microbe-drug-disease tripartite network are shown in Fig. 3.

In the course of the experiment, we use two layers of RGCN ($l = 2$), to learn node features in the microbe-drug-disease tripartite network. After that, we get the prediction score by calculating the node features of each microbe and the disease with dot product.

We use Adam Optimizer [31] to train the models by optimizing the cross entropy loss function. The formula of cross-entropy

TABLE II
METHOD PERFORMANCE OF DIFFERENT DIMENSIONS

Hidden 1	Hidden 2	AUC	AUPR	Precision	Recall
128	128	0.9030	0.8904	0.8127	0.6732
128	64	0.9038	0.8914	0.8142	0.6722
128	32	0.9020	0.8878	0.808	0.6775
64	64	0.9034	0.8920	0.8129	0.6746
64	32	0.9007	0.8834	0.8146	0.6653

loss is shown as (13):

$$Loss = \sum_{(d_i, m_j) \in E} -label(d_i, m_j) \log(p(d_i, m_j)) - (1 - label(d_i, m_j)) \log(1 - p(d_i, m_j)) \quad (13)$$

where E contains all edges where microbes and diseases are connected. $label(d_i, m_j)$ is the real label of the connection between disease i and microbe j . $p(d_i, m_j)$ is the predicted score between disease i and microbe j .

III. EXPERIMENTS AND RESULTS

In order to evaluate the prediction performance of TNRGCN, we implement 5-fold cross validation. We randomly divide all microbe-disease associations into five groups. Each group is treated as test samples, while others are training samples. In this validation, every confirmed microbe-disease association is regarded as positive sample. In order to balance the positive and negative samples, we randomly select unconfirmed microbe-disease associations equal to positive samples as negative samples. In order to avoid the result deviation caused by different sample segmentation, we run the cross validation for 10 times and average the scores. According to predicted scores, we plot receiver operating characteristic (ROC) curve and precision-recall curve, and the performance of the method is evaluated by the area under the receiver operating characteristic (AUC) and the area of precision recall curve (AUPR).

A. Parameter Selection

We consider the feature dimensions of two layers to select the best parameter combination for this method. We take different dimensions between 32 and 128 to analyze the parameters. According to Table II, the dimensions of feature are set to 128 and 64 in the first layer and the second layer, respectively.

B. Model Analysis

In order to evaluate the influence of different processes in this method, we compare TNRGCN with its different variants, which are as follows:

TNRGCNRF: it initializes features randomly.

BNRGCN: it utilizes two-layer RGCN to predict microbe-disease associations on microbe-disease bipartite network, without drugs.

ONRGCN: it utilizes one-layer RGCN to predict microbe-disease associations on microbe-drug-disease tripartite network.

TABLE III
PREDICTION PERFORMANCE OF TNRGCN AND ITS VARIANTS

Variants	AUC	AUPR	Precision	Recall
TNRGCN	0.9038	0.8914	0.8142	0.6722
TNRGCNRF	0.8232	0.8090	0.7355	0.6625
BNRGCN	0.8938	0.8803	0.7922	0.6857
ONRGCN	0.7683	0.7902	0.7216	0.6301
TNGCN	0.6847	0.7186	0.6728	0.5892

TABLE IV
DETAILS OF HMDAD AND DISBIOME

Database	Microbe	Disease	Associations	AUC	AUPR
HMDAD	292	39	450	0.9332	0.9030
Disbiome	1416	243	7052	0.8926	0.8678

TNGCN: it utilizes two-layer GCN to predict microbe-disease associations on microbe-drug-disease tripartite network.

As shown in Table III, the predictive performance is higher when it uses similarities as initial features and adds drug nodes. Compared with one-layer RGCN, two-layer RGCN can achieve better performance. This is because only the first-order neighbor information of the node is aggregated by using one-layer RGCN. Through two layers RGCN, nodes can learn more neighbor information. Thus, we can see that the indirect connection between nodes has an important impact on the prediction performance.

Besides, we compare the impact of using RGCN and GCN on prediction performance. From the results, it can be seen that the prediction results of RGCN are significantly higher than that of GCN. The low prediction performance of GCN may be due to the fact that the method does not consider the type of edges, and the connection between microbe-drug and disease-drugs leads to biased prediction weights. Instead, RGCN considers different types of edge weights, and is more suitable for networks with multiple types of nodes in reality.

In order to evaluate the adaptive performance of TNRGCN, we used five-fold cross-validation on HMDAD and Disbiome, respectively. The datasets and results are shown in Table IV.

According to Table IV, we can see that TNRGCN is equally suitable for prediction on small datasets. Its AUC value on HMDAD reaches 0.9332. We believe that this is because the HMDAD is small, so it has relatively more associations and better prediction results.

C. Comparison With Other Methods

We compare TNRGCN with nine methods under 5-fold cross validation. These methods include BRWMDA [32], BDSILP [29], BiRWHMDA [33], KATZHMDA [11], NTSHMDA [34], KATZBNRA [12], NCPHMDA [35], PBHMDA [36], and NC-PLP [37]. According to Fig. 4 and Table V, the AUC and AUPR values of TNRGCN are both the highest.

By comparison, we can see that most baseline methods perform poorly. We believe that this is due to the sparseness of the huge biological network. Most methods pass information through matrix operations, resulting in loss of information. RGCN only considers edges and transmits information through

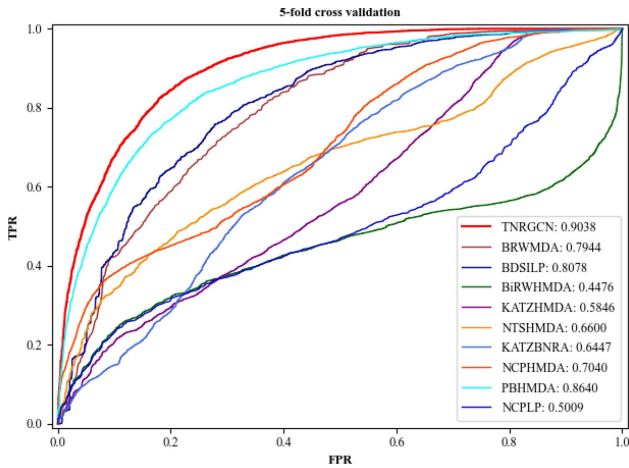


Fig. 4. ROC curve of ten methods in 5-fold cross validation.

TABLE V

AUPR PREDICTED BY TEN METHODS IN 5-FOLD CROSS VALIDATION

Methods	AUPR	Methods	AUPR
TNRGCN	0.8914	NTSHMDA	0.6828
BRWMDA	0.7469	KATZBNRA	0.6061
BDSILP	0.7564	NCPHMDA	0.7225
BiRWHMDA	0.5593	PBHMDA	0.8577
KATZHMDA	0.5819	NCPLP	0.5722

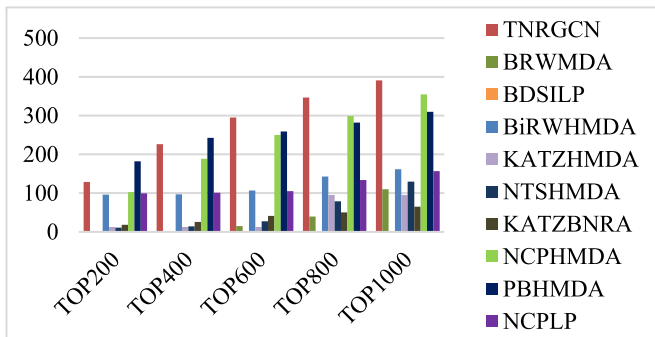


Fig. 5. The number of correctly associations predicted by ten methods in 5-fold cross validation.

edges, which effectively avoids the situation of data loss caused by matrix operations.

Then, we compare the prediction scores of these methods in the 1000 microbe-disease pairs with the highest scores. As shown in Fig. 5, the number of confirmed microbe-diseases associations predicted by is more among ten methods.

We also calculate the AUC value for each individual disease. Fig. 6 shows that AUCs for most diseases predicted by TNRGCN are above 0.8, with the higher mean and median value compared with others. As shown in Table VI, we performed t-hypothesis tests on different methods to test the degree of difference in the mean values of TNRGCN and other methods. The p -values are all less than 0.05, indicating that TNRGCN is significantly different from other methods.

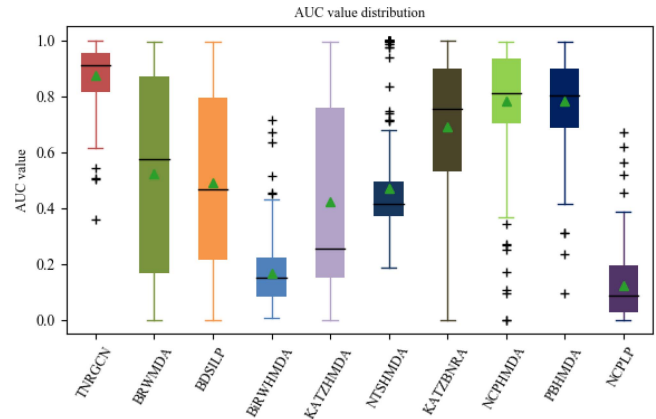


Fig. 6. Distribution of AUC value for all diseases in 5-fold cross validation.

TABLE VI

PAIRWISE T-HYPOTHESIS TESTING BETWEEN TNRGCN AND OTHER METHODS

Methods	p-value	Methods	p-value
BRWMDA	4.0292e-49	KATZBNRA	1.0989e-37
BDSILP	2.0949e-64	NCPHMDA	2.6303e-24
BiRWHMDA	3.7234e-148	PBHMDA	1.1924e-29
KATZHMDA	1.9864e-71	NCPLP	8.3705e-145
NTSHMDA	7.5798e-106		

TABLE VII

TOP-10 MICROBES ASSOCIATED WITH T2D

Rank	Microbe	Evidence
1	Desulfovibrio	PMID: 23023125
2	Ruminococcus	PMID: 32490229
3	Rothia	PMID: 28648853
4	Parvimonas	unconfirmed
5	Porphyromonas	PMID: 22762355
6	Prevotella melaninogenica	PMID: 33930650
7	Anaerotruncus	PMID: 32952631
8	Enterobacteriaceae	PMID: 32694777
9	Peptostreptococcaceae	unconfirmed
10	Barnesiella	PMID: 33990912

IV. CASE STUDIES

We implement case studies for T2D, Bipolar disorder and Autism to evaluate the predictive performance of TNRGCN. After sorting scores of unconfirmed microbe-disease associations in descending order, we select top 10 microbes corresponding to each disease.

T2D is a chronic disease due to impaired insulin secretion, which is accompanied by a series of health problems such as kidney failure, cardiovascular disease and weakness [38]. In this article, we select T2D for case study. As shown in Table VII, VIII out of top 10 potential microbes are confirmed. For example, Ruminococcus can be regarded as taxonomic biomarkers for elderly diabetic patients [39]. Compared with nondiabetic individuals, the percentages of Porphyromonas and Prevotella melaninogenica are lower, while Desulfovibrio is enriched in T2D patients [40], [41], [42]. In T2D patients, the

TABLE VIII
TOP-10 MICROBES ASSOCIATED WITH BIPOLAR DISORDER

Rank	Microbe	Evidence
1	Streptococcus	PMID: 30927646
2	Lactobacillus	PMID: 32718072
3	Prevotella	Unconfirmed
4	Bifidobacterium	PMID: 30927646
5	Blautia	Unconfirmed
6	Veillonella	Unconfirmed
7	Enterococcus	unconfirmed
8	Lachnospiraceae	PMID: 32274300
9	Enterobacteriaceae	PMID: 30838027
10	Corynebacterium	unconfirmed

proportion of *Rothia* increases dramatically in oral microbiota [43], the level of Gram-negative Enterobacteriaceae is enhanced in mesenteric adipose [44]. Moreover, in the intestines of T2D patients, the abundance of *Anaerotruncus* is lower, which is increasing slightly during treatment [45]. What's more, in the analysis of T2D-associated gut microbiota, *Barnesiella* is found to be associated with the incidence of Mongolian T2D [46].

Bipolar disorder includes bipolar I disorder and bipolar II disorder, which is often characterized by manic episode or depression and usually occurs in adolescence [47]. At present, the pathogenesis of bipolar disorder is still unclear. In the case study of bipolar disorder, 5 of the top 10 potential microbes are confirmed, which is shown in Table VIII. For instance, through comparing the differences of intestinal microbiota with healthy participants, the number of *Streptococcus* and *Bifidobacterium* increase significantly in the bipolar disorder participants at genus level [48]. By examining bacterial counts in fecal samples from patients with bipolar disorder, studies found that there is a negative correlation between *Lactobacillus* counts and sleep, which is beneficial to the improvement of sleep quality [49]. Research also found that the abundance of Enterobacteriaceae is increased in bipolar disorder's patients [50]. Although *Prevotella* and *Blautia* are not proved to be directly related to bipolar disorder, researches have found that they were all associated with mental or behavioral dysfunction such as major depressive disorder and schizophrenia [51], [52].

Autism is a generalized developmental disorder that combines cognitive function, language function, and interpersonal social communication with special pathologies, resulting in significant difficulties in adapting to social life. Studies have demonstrated a strong positive relationship between autism severity and gastrointestinal dysfunction severity [53]. The combination of diet and probiotics to modulate the composition of gut microbiota is beneficial for the treatment of children with autism [54], [55]. Thus, we do a case study on autism. As shown in Table IX, VIII of the top 10 potential microbes predicted by TNRGCN are confirmed. For example, in studies of microbiome profiles autistic patients and controls, the relative abundance of *Dialister*, *Parabacteroides* in autistic patients are lower than neurotypical controls, while *Klebsiella* is significantly higher [56], [57]. A study for fecal microbiota of children with autism and healthy children found that at the genus level, the number of

TABLE IX
TOP-10 MICROBES ASSOCIATED WITH AUTISM

Rank	Microbe	Evidence
1	<i>Pseudomonas</i>	unconfirmed
2	<i>Dialister</i>	PMID: 34095558
3	<i>Haemophilus</i>	PMID: 32830918
4	<i>Leptotrichia</i>	unconfirmed
5	<i>Klebsiella</i>	PMID: 34095558
6	<i>Actinomyces</i>	PMID: 29371629
7	Lachnospiraceae	PMID: 30071894
8	Proteobacteria	PMID: 34487668
9	Ruminococcus	PMID: 34753541
10	<i>Parabacteroides</i>	PMID: 31404299

Haemophilus bacteria is reduced in children with autism compared with controls [58]. In addition, through high-throughput sequencing of oral samples, the bacterial diversity observed in children with autism was lower compared to controls. The abundance of *Actinomyces* in the patient's saliva was reduced [59].

V. CONCLUSION

Knowing associations between microbes and diseases is beneficial to disease diagnosis and treatment. In this article, we propose a method for microbe-disease association prediction called TNRGCN based on Tripartite Network and RGCN. First, considering that HMDAD contains relatively few microbes, diseases and associations, we integrate HMDAD and Disbiome to obtain more related information. More associations can improve the performance of predictive method. Second, we introduce the drug information from MDAD and CTD to increase the indirect associations in the microbe-disease network, thus building a microbe-drug-disease tripartite network. Third, we utilize RGCN on the microbe-drug-disease tripartite network to predict potential microbe-disease associations, which can be applied to heterogeneous networks containing different types of nodes and edges. TNRGCN has a good performance in 5-fold cross validation. Experiments of case studies further demonstrate the predictive performance of TNRGCN.

However, the number of microbe-drug associations and disease-drug associations involved is relatively small although we have added drug related associations in data process. With the establishment of more databases, we will integrate more confirmed associations in future work. In addition, with the in-depth study of multi-omics such as genomics [60], proteomics [61] and transcriptomics [62], the combination of multi-omics data may further excavate more biological information, which is conducive to the prediction, diagnosis and treatment of diseases [63].

REFERENCES

- [1] H. M. P. Consortium, "A framework for human microbiome research," *Nature*, vol. 486, no. 7402, pp. 215–221, Jun. 14, 2012.
- [2] L. Gao et al., "Oral microbiomes: More and more importance in oral cavity and whole body," *Protein Cell*, vol. 9, no. 5, pp. 488–500, May 2018.

- [3] T. G. Dinan et al., "Collective unconscious: How gut microbes shape human behavior," *J. Psychiatr. Res.*, vol. 63, pp. 1–9, Apr. 2015.
- [4] V. B. Young, "The role of the microbiome in human health and disease: An introduction for clinicians," *Brit. Med. J.*, vol. 356, no. j831, pp. 1–14, Mar. 2017.
- [5] L. Dethlefsen, M. McFall-Ngai, and D. A. Relman, "An ecological and evolutionary perspective on human-microbe mutualism and disease," *Nature*, vol. 449, no. 7164, pp. 811–818, Oct. 2007.
- [6] K. Makki et al., "The Impact of dietary fiber on gut microbiota in host health and disease," *Cell Host Microbe*, vol. 23, no. 6, pp. 705–715, Jun. 13, 2018.
- [7] J. D. Forbes, G. Van Domselaar, and C. N. Bernstein, "The gut microbiota in immune-mediated inflammatory diseases," *Front. Microbiol.*, vol. 7, no. 1081, pp. 1–18, Jul. 11, 2016.
- [8] D. Kashtanova et al., "Gut microbiota and vascular biomarkers in patients without clinical cardiovascular diseases," *Artery Res.*, vol. 18, pp. 41–48, Jun. 2017.
- [9] W. Ma et al., "An analysis of human microbe-disease associations," *Brief. Bioinf.*, vol. 18, no. 1, pp. 85–97, Jan. 2017.
- [10] Y. Janssens et al., "Disbiome database: Linking the microbiome to disease," *BMC Microbiol.*, vol. 18, no. 50, pp. 1–6, Jun. 4, 2018.
- [11] X. Chen et al., "A novel approach based on KATZ measure to predict associations of human microbiota with non-infectious diseases," *Bioinformatics*, vol. 33, no. 5, pp. 733–739, Mar. 1, 2017.
- [12] H. Li et al., "A novel human microbe-disease association prediction method based on the bidirectional weighted network," *Front. Microbiol.*, vol. 10, no. 676, pp. 1–13, Apr. 9, 2019.
- [13] C. Jiang, M. Tang, J. Shuting, W. Huang, and X. Liu, "KGNMDA: A knowledge graph neural network method for predicting microbe-disease associations," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, early access, Jun. 20, 2022, doi: [10.1109/TCBB.2022.3184362](https://doi.org/10.1109/TCBB.2022.3184362).
- [14] L. Peng et al., "Prioritizing human microbe-disease associations utilizing a node-information-based link propagation method," *IEEE Access*, vol. 8, pp. 31341–31349, 2020.
- [15] M. F. Hua et al., "MVGCNMDA: Multi-view graph augmentation convolutional network for uncovering disease-related microbes," *Interdiscipl. Sci.-Comput. Life Sci.*, vol. 14, no. 3, pp. 669–682, Sep. 2022.
- [16] H. Li et al., "Identifying microbe-disease association based on a novel back-propagation neural network model," *IEEE-ACM Trans. Comput. Biol. Bioinf.*, vol. 18, no. 6, pp. 2502–2513, Nov. 1, 2021.
- [17] Y. Y. Wang, X. J. Lei, C. Lu, and Y. Pan, "Predicting microbe-disease association based on multiple similarities and LINE algorithm," *IEEE-ACM Trans. Comput. Biol. Bioinf.*, vol. 19, no. 4, pp. 2399–2408, Jul./Aug. 2022.
- [18] L. Dayun, L. Junyi, L. Yi, H. Qihua, and L. Deng, "MGATMDA: Predicting microbe-disease associations via multi-component graph attention network," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 19, no. 6, pp. 3578–3585, Sep. 2021, doi: [10.1109/TCBB.2021.3116318](https://doi.org/10.1109/TCBB.2021.3116318).
- [19] W. Peng, M. Liu, W. Dai, T. Chen, Y. Fu, and Y. Pan, "Multi-view feature aggregation for predicting microbe-disease association," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, early access, Dec. 06, 2021, doi: [10.1109/TCBB.2021.3132611](https://doi.org/10.1109/TCBB.2021.3132611).
- [20] Y. L. Chen and X. J. Lei, "Metapath aggregated graph neural network and tripartite heterogeneous networks for microbe-disease prediction," *Front. Microbiol.*, vol. 13, May 2022, Art. no. 919380.
- [21] T. N. Kip F and M. Welling, "Semi-supervised classification with graph convolutional networks," 2016, *arXiv:1609.02907*.
- [22] T. B. Mudiyansele et al., "Predicting CircRNA disease associations using novel node classification and link prediction models on graph convolutional networks," *Methods*, vol. 198, pp. 32–44, Feb. 2022.
- [23] X. J. Lei, J. J. Tie, and Y. Pan, "Inferring metabolite-disease association using graph convolutional networks," *IEEE-ACM Trans. Comput. Biol. Bioinf.*, vol. 19, no. 2, pp. 688–698, Mar./Apr. 2022.
- [24] F. Wang et al., "Predicting drug-drug interactions by graph convolutional network with multi-kernel," *Brief. Bioinf.*, vol. 23, no. 1, Jan. 17, 2022, Art. no. bbab511.
- [25] M. Schlichtkrull et al., "Modeling relational data with graph convolutional networks," in *Proc. Eur. Semantic Web Conf.*, 2018, pp. 593–607.
- [26] J. Pinero et al., "DisGeNET: A discovery platform for the dynamical exploration of human diseases and their genes," *Database*, vol. 2015, 2015, Art. no. bav028.
- [27] C. Yan, G. Duan, F.-X. Wu, Y. Pan, and J. Wang, "MCHMDA: Predicting microbe-disease associations based on similarities and low-rank matrix completion," *IEEE-ACM Trans. Comput. Biol. Bioinf.*, vol. 18, no. 2, pp. 611–620, Mar./Apr. 2021.
- [28] I. K. Dhammi and S. Kumar, "Medical subject headings (MeSH) terms," *Indian J. Orthopaedics*, vol. 48, no. 5, pp. 443–444, Sep./Oct. 2014.
- [29] W. Zhang, W. T. Yang, X. T. Lu, F. Huang, and F. Luo, "The bi-direction similarity integration method for predicting microbe-disease associations," *IEEE Access*, vol. 6, pp. 38052–38061, 2018.
- [30] T. van Laarhoven, S. B. Nabuurs, and E. Marchiori, "Gaussian interaction profile kernels for predicting drug-target interaction," *Bioinformatics*, vol. 27, no. 21, pp. 3036–3043, Nov. 1, 2011.
- [31] D. Kingma and J. J. C. S. Ba, "Adam: A method for stochastic optimization," Dec. 22, 2014, *arXiv:1412.6980*.
- [32] C. Yan, G. H. Duan, F. X. Wu, Y. Pan, and J. Wang, "BRWMDA: Predicting microbe-disease associations based on similarities and bi-random walk on disease and microbe networks," *IEEE-ACM Trans. Comput. Biol. Bioinf.*, vol. 17, no. 5, pp. 1595–1604, Sep. 1, 2020.
- [33] S. Zou, J. P. Zhang, and Z. P. Zhang, "A novel approach for predicting microbe-disease associations by bi-random walk on the heterogeneous network," *Plos One*, vol. 12, no. 9, Sep. 7, 2017, Art. no. e0184394.
- [34] J. W. Luo and Y. H. Long, "NTSHMDA: Prediction of human microbe-disease association based on random walk by integrating network topological similarity," *IEEE-ACM Trans. Comput. Biol. Bioinf.*, vol. 17, no. 4, pp. 1341–1351, Jul./Aug. 2020.
- [35] W. Z. Bao, Z. C. Jiang, and D. S. Huang, "Novel human microbe-disease association prediction using network consistency projection," *BMC Bioinf.*, vol. 18, no. 543, pp. 173–181, Dec. 28, 2017.
- [36] Z. A. Huang et al., "PBHMDA: Path-based human microbe-disease association prediction," *Front. Microbiol.*, vol. 8, Feb. 22, 2017, Art. no. 233.
- [37] M. M. Yin, J. X. Liu, Y. L. Gao, X.-Z. Kong, and C.-H. Zheng, "NCPLP: A novel approach for predicting microbe-associated diseases with network consistency projection and label propagation," *IEEE Trans. Cybern.*, vol. 52, no. 6, pp. 5079–5087, Jun. 2022.
- [38] H. C. Gerstein et al., "Effects of intensive glucose lowering in type 2 diabetes," *New England J. Med.*, vol. 358, no. 24, pp. 2545–2559, Jun. 12, 2008.
- [39] A. O. Afolayan et al., "Insights into the gut microbiota of Nigerian elderly with type 2 diabetes and non-diabetic elderly persons," *Heliyon*, vol. 6, no. 5, May 2020, Art. no. e03971.
- [40] R. C. V. Casarin et al., "Subgingival biodiversity in subjects with uncontrolled type-2 diabetes and chronic periodontitis," *J. Periodontal Res.*, vol. 48, no. 1, pp. 30–36, Feb. 2013.
- [41] Y. K. Liu et al., "A salivary microbiome-based auxiliary diagnostic model for type 2 diabetes mellitus," *Arch. Oral Biol.*, vol. 126, Jun. 2021, Art. no. 105118.
- [42] J. J. Qin et al., "A metagenome-wide association study of gut microbiota in type 2 diabetes," *Nature*, vol. 490, pp. 55–60, Oct. 4, 2012.
- [43] R. Anbalagan et al., "Next generation sequencing of oral microbiota in Type 2 diabetes mellitus prior to and after neem stick usage and correlation with serum monocyte chemoattractant-1," *Diabetes Res. Clin. Pract.*, vol. 130, pp. 204–210, Aug. 2017.
- [44] F. F. Anhe et al., "Type 2 diabetes influences bacterial tissue compartmentalisation in human obesity," *Nature Metab.*, vol. 2, no. 3, pp. 233–242, Mar. 2020.
- [45] M. Zhang et al., "The gut microbiome can be used to predict the gastrointestinal response and efficacy of lung cancer patients undergoing chemotherapy," *Ann. Palliat. Med.*, vol. 9, no. 6, pp. 4211–4227, Nov. 2020.
- [46] S. C. Li et al., "Comparative analysis of type 2 diabetes-associated gut microbiota between Han and Mongolian people," *J. Microbiol.*, vol. 59, no. 7, pp. 693–701, Jul. 2021.
- [47] R. S. McIntyre et al., "Bipolar disorders," *Lancet*, vol. 396, no. 10265, pp. 1841–1856, Dec. 2020.
- [48] H. Rong et al., "Similarly in depression, nuances of gut microbiota: Evidences from a shotgun metagenomics sequencing study on major depressive disorder versus bipolar disorder with current major depressive episode patients," *J. Psychiatr. Res.*, vol. 113, pp. 90–99, Jun. 2019.
- [49] E. Aizawa et al., "Bifidobacterium and lactobacillus counts in the gut microbiota of patients with bipolar disorder and healthy controls," *Front. Psychiatry*, vol. 9, no. 730, pp. 1–8, Jan. 2019.
- [50] T. T. Huang et al., "Current understanding of gut microbiota in mood disorders: An update of human studies," *Front. Genet.*, vol. 10, no. 98, pp. 1–12, Feb. 2019.
- [51] McGuinness A. J. et al., "A systematic review of gut microbiota composition in observational studies of major depressive disorder, bipolar disorder and schizophrenia," *Mol. Psychiatry*, vol. 27, no. 4, pp. 1920–1935, Apr. 2022.

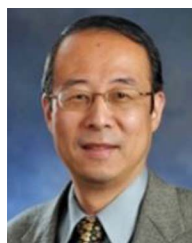
- [52] K. Yamaoka, N. Uotsu, and E. Hoshino, "Relationship between psychosocial stress-induced prefrontal cortex activity and gut microbiota in healthy Participants-A functional near-infrared spectroscopy study," *Neurobiol. Stress*, vol. 20, no. 100479, pp. 1–14, Sep. 2022.
- [53] A. Tomova et al., "Gastrointestinal microbiota in children with autism in Slovakia," *Physiol. Behav.*, vol. 138, pp. 179–187, Jan. 2015.
- [54] R. Grimaldi et al., "A prebiotic intervention study in children with autism spectrum disorders (ASDs)," *Microbiome*, vol. 6, no. 1, Aug. 2, 2018, Art. no. 133.
- [55] Y.-Q. Li et al., "Effect of probiotics combined with applied behavior analysis in the treatment of children with autism spectrum disorder: A prospective randomized controlled trial," *Randomized Controlled Trial*, vol. 23, no. 11, pp. 1103–1110, Nov. 2021.
- [56] F. Ye et al., "Comparison of gut microbiota in autism spectrum disorders and neurotypical boys in China: A case-control study," *Synthetic Syst. Biotechnol.*, vol. 6, no. 2, pp. 120–126, Jun. 2021.
- [57] M. Y. Xu et al., "Association between gut microbiota and autism spectrum disorder: A systematic review and meta-analysis," *Front. Psychiatry*, vol. 10, no. 473, pp. 1–11, Jul. 17, 2019.
- [58] R. Zou et al., "Changes in the gut microbiota of children with autism spectrum disorder," *Autism Res.*, vol. 13, no. 9, pp. 1614–1625, Sep. 2020.
- [59] Y. A. Qiao et al., "Alterations of oral microbiota distinguish children with autism spectrum disorders from healthy controls," *Sci. Rep.s*, vol. 8, no. 1597, pp. 1–12, Jan. 25, 2018.
- [60] P. Mobadersany et al., "Predicting cancer outcomes from histology and genomics using convolutional networks," *Proc. Nat. Acad. Sci. USA*, vol. 115, no. 13, pp. E2970–E2979, Mar. 2018.
- [61] M. Zeng et al., "A deep learning framework for identifying essential proteins by integrating multiple types of biological information," *IEEE-ACM Trans. Comput. Biol. Bioinf.*, vol. 18, no. 1, pp. 296–305, Jan. 2021.
- [62] X. J. Lei et al., "A comprehensive survey on computational methods of non-coding RNA and disease association prediction," *Brief. Bioinf.*, vol. 22, no. 4, Jul. 2021, Art. no. bbaa350.
- [63] Y. Pan, X. J. Lei, and Y. C. Zhang, "Association predictions of genomics, proteomics, transcriptomics, microbiome, metabolomics, pathomics, radiomics, drug, symptoms, environment factor, and disease networks: A comprehensive approach," *Med. Res. Rev.*, vol. 42, no. 1, pp. 441–461, Jan. 2022.



Yueyue Wang received the BS degree from the School of Computer Science, Shaanxi Normal University, Xi'an, China, in 2019, where she is currently working toward the MS degree. Her current research interests include bioinformatics, data mining, and deep learning.



Xiujuan Lei received the MS and PhD degrees from Northwestern Polytechnical University, Xi'an, China, in 2001 and 2005, respectively. She is currently a professor with the School of Computer Science, Shaanxi Normal University, Xi'an. Her research interests include bioinformatics, swarm intelligent optimization, data mining, and deep learning.



Yi Pan received the BEng and MEng degrees in computer engineering from Tsinghua University, China, in 1982 and 1984, respectively, and the PhD degree in computer science from the University of Pittsburgh, USA, in 1991. He is currently a professor with the Faculty of Computer Science and Control Engineering, Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences. He has served as chair of Computer Science Department, Georgia State University during 2005–2020. His current research interests mainly include bioinformatics and health informatics using Big Data analytics, cloud computing, and machine learning technologies.