KGNMDA: A Knowledge Graph Neural Network Method for Predicting Microbe-Disease Associations

Changzhi Jiang, Minli Tang, Shuting Jin, Wei Huang, and Xiangrong Liu

Abstract—Accumulated studies discovered that various microbes in human bodies were closely related to complex human diseases and could provide new insight into drug development. Multiple computational methods were constructed to predict microbes that were potentially associated with diseases. However, most previous methods were based on single characteristics of microbes or diseases, that lacked important biological information related to microorganisms or diseases. Therefore, we constructed a knowledge graph centered on microorganisms and diseases from several existed databases to provide knowledgeable information for microbes and diseases. Then, we adopted a graph neural network method to learn representations of microbes and diseases from the constructed knowledge graph. After that, we introduced the Gaussian kernel similarity features of microbes and diseases to generate final representations of microbes and diseases. At last, we proposed a score function on final representations of microbes and diseases to predict scores of microbe-disease associations. Comprehensive experiments on the Human Microbe-Disease Association Database (HMDAD) dataset had demonstrated that our approach outperformed baseline methods. Furthermore, we implemented case studies on two important diseases (asthma and inflammatory bowel disease), the result demonstrated that our proposed model was effective in revealing the relationship between diseases and microbes. The source code of our model and the data were available on https://github.com/ChangzhiJiang/KGNMDA_master.

Index Terms—Microbe-disease associations, knowledge graph, graph neural network

1 Introduction

Microorganism or microbe exists in its single-celled form or a colony of cells, which is composed of bacteria, archaea, fungi, viruses, and protozoa [1]. Microbes that distribute on human skin, respiratory tract, oral cavity, gastrointestinal tract, and other parts are important for human health [2]. For example, the intestinal microbiota plays a critical role in the maturation and continued education of the host immune response [3]. The abnormality of the microbial communities could lead to diseases, such as Inflammatory bowel disease (IBD) [4], asthma [5], and even cancer [6]. Therefore, it is important that understands the roles of microbes in the pathogenic mechanism of diseases.

Though recovering the microbe-disease relations could assist to analyze the complex pathogenesis of diseases and provide novel insights for developing drugs, the conventional experimental methods were time-consuming, expensive, and laborious. Therefore, accumulated researches designed computational methods to help identify microbe-disease associations. For instance, a network-based KATZ model was proposed to recover microbe-disease associations [7]. Long *et al.* [8] designed a meta-graph pattern to infer candidate microbe-disease pairs on a heterogeneous network. Zou *et al.* [9] introduced a bi-random walk on the heterogeneous network for predicting microbe-disease relations. Besides, a novel improved random walk-based method in which integrated network topological similarity was proposed by Luo *et al.* [10]. In addition, matrix factorization or completion methods were also introduced to solve this task. For example, a collaborative matrix factorization model was designed to predict microbe-disease associations [11]. He *et al.* [12] proposed a predictive model of graph regularized non-negative matrix factorization to reveal microbe-disease associations. Yan *et al.* [13] developed a low-rank matrix completion model to do that.

The aforementioned methods for predicting microbe-disease associations performed well. However, those methods still had limitations. For instance, network-based methods have shown poor performance in diseases or microbes which had few known associations. Because there was a difference in information propagation between a sparse and dense part of the network. Matrix factorization (or completion)-based can not learn nonlinear information in the microbe-disease pairs. Therefore, we considered that added more knowledge

- *Changzhi Jiang and Wei Huang are with the School of Informatics, Xiamen University, Xiamen 361005, China. E-mail: {czjiang, hungwii}@stu.xmu.edu.cn.*
- *Minli Tang is with the School of Informatics, Xiamen University, Xiamen 361005, China, and also with the School of Big Data Engineering, KaiLi University, KaiLi 556011, China. E-mail: tangml@stu.xmu.edu.cn.*
- *Shuting Jin is with the School of Informatics, National Institute for Data Science in Health and Medicine, Xiamen University, Xiamen 361005, China. E-mail: stjin@stu.xmu.edu.cn.*
- *Xiangrong Liu is with the School of Informatics, National Institute for Data Science in Health and Medicine, National and Local Joint Engineering Research Center for Navigation and Location Service Technology, Xiamen University, Xiamen 361005, China. E-mail: xrliu@xmu.edu.cn.*

Manuscript received 17 Aug. 2021; revised 20 May 2022; accepted 1 June 2022. Date of publication 20 June 2022; date of current version 3 Apr. 2023.
This work was supported in part by the National Natural Science Foundation of China under Grants 62072384, 61872309, 62072385, and 61772441 and in part by the Zhejiang Lab under Grant 2022RD0AB02.
(Corresponding author: Xiangrong Liu.)
Digital Object Identifier no. 10.1109/TCBB.2022.3184362

information to solve the limitations. Recently, knowledge graph (KG) which was a form of structured human knowledge had drawn great attention from both academia and the industry [15]. In the KG, the nodes represented the entities, the edges represented the relations between entities, and facts were modeled as three triples. For example, the triples (1,3-Dimethyluric acid,metabolite-disease, Inflammatory bowel disease) in KG can represent that the metabolite 1,3-Dimethyluric acid was associated with the Inflammatory bowel disease [16]. That could provide a wealth of knowledge for the Inflammatory bowel disease. There had been several studies that introduced KG for link prediction in the bioinformatics domain and made significant progress, such as the drug-target interactions prediction task [17] and the drug-drug interaction prediction problem [20]. Furthermore, various algorithms were proposed to learn KG representations e.g., TransE [18] and graph neural network [19]. For example, Lin *et al.* [20] adopted a graph neural network algorithm to learn KG embedding to predict the drug-drug interaction.

Inspired from the work of Lin *et al.* [20], we proposed a novel knowledge graph and graph neural network-based method to predict microbe-disease associations (KGNMDA). At first, we constructed a KG centered on microorganisms and diseases from several existed databases. Then, we adopted a graph neural network (GNN) method to learn representations of microbes and diseases from our constructed KG. After that, we combined Gaussian kernel similarity features of microbes (or diseases) and representations learned from the previous step to generate the final feature vectors of microbes and diseases. At last, we proposed a score function to predict the scores of microbe-disease associations using the final feature vectors of microbes and diseases. We conducted comprehensive experiments to evaluate the performance of our proposed KGNMDA model. The experimental result demonstrated our proposed model outperformed the compared methods. Furthermore, we implemented case studies on two important diseases, i.e., asthma and inflammatory bowel disease, that further shown our proposed model was effective for identifying microbe-disease associations. In summary, the contributions of our work were as follows:

a) We constructed a KG centered on microorganisms and diseases from several existed databases. That could provide knowledge information of microbes and diseases for research about microbes or diseases in the future.
b) This was the first attempt to construct KG on microbe-disease associations prediction task.
c) We proposed a KGNMDA model to predict scores of microbe-disease associations and experiments demonstrated KGNMDA was effective in revealing the relationship between diseases and microbes.

2 Materials

2.1 Human Microbe-Disease Associations

We downloaded the known experimentally validated human microbe-disease associations data from Human Microbe-Disease Association Database (http://www.cuilab.cn/hmdad), which contained 483 experimentally confirmed microbe-disease associations including 39 diseases and 292 microbes [21]. In HMDAD, we found that a microbe-disease pair may include several entries from different evidence. This would lead to the problem of data redundancy. Therefore, we only kept a microbe-disease pair when the same microbe-disease associations from different evidence. Finally, we gained 449 microbe-disease associations involving 39 diseases and 291 microbes.

TABLE 1
The Details of the Entity and Relation in the MDKG

MDKG	Entity	Entity-type	Triplet	Relation-type
Count	66892	9	316831	39

2.2 Construct Knowledge Graph Involving Diseases and Microbes

To gain informative knowledge from biomedical databases for diseases and microbes, we constructed a knowledge graph named MDKG (Microbe Disease Knowledge Graph) for microbe-disease associations prediction. The MDKG contained biomedical entities that were extracted from different data sources including DRKG [22], VMH [23], HMDB [24], MIND (http://microbialnet.org/mind.html), and NCBI [25]. We would keep a unified name for the same biological entity that was from different data sources. That could remove redundant data to improve quality for our MDKG. MDKG included 66,892 entities belonging to 9 entity-types, and 316,831 triplets belonging to 39 edge-types. These 39 edge-types showed a type of interaction between one of the 11 entity-type pairs (multiple types of interactions were possible between the same entity-pair). The details of the MDKG were demonstrated in Tables 1, 2, and 3. Note, our MDKG was centered on microorganisms and diseases. For instance, microbe-microbe associations provided the relation information between different microbes. Metabolite-disease pairs showed the diseases were related to metabolites in our bodies. There was biomedical knowledge about microbes generated metabolites in the microbe-metabolite associations.

3 Methods

In this section, we first formulated the microbe-disease associations prediction problem. Then, we would introduce the KGNMDA model in detail. As shown in Fig. 1, KGNMDA included three main steps. At first, we learned feature representations of microbes and diseases from MDKG based on GNN. Then, we combined the Gaussian similarity features of microbes and diseases and representations learned from the previous step. After that, the informative embeddings of microbes and diseases were obtained. At last, we implemented a score function on final embeddings of microbes and diseases to output scores of microbe-disease associations.

3.1 Problem Formulation

We formatted an adjacency matrix $A \in R^{nd \times nm}$ that represented the known human microbe-disease associations, where nd and nm denoted the number of diseases and

TABLE 2
The Details of the Entity-Type in the MDKG

Entity-type	Count	Entity-type	Count	Entity-type	Count
Microbe	5179	Disease	5626	Metabolite	23000
Compound	9572	Organ	55	Gene	22536
Anatomy	398	Division	111	Symptom	415

microbes, separately. If there existed the experimentally confirmed association between disease d and microbe m, we set $A_{(d,m)} = 1$, otherwise $A_{(d,m)} = 0$. In addition, we denoted MDKG by $G = (E, R)$, which included entity-relation-entity triples, where E and R was the set of entities and relations, respectively. To a triple $T = (h, r, t)$, where $h, t \in E$ and $r \in R$, that represented a relationship of type r between the head entity h and the tail entity t in the triple. Given the matrix A and MDKG G, we aimed to predict the score between a microbe m and a disease d. In summary, we hoped to learn a prediction function $\hat{y}_{d,m} = \Phi(d, m \,|\, \alpha, A, G)$, where $\hat{y}_{d,m}$ was the score of prediction between a microbe m and a disease d, and α represented the parameters of function Φ.

3.2 Learning MDKG Representations

At first, we introduced the part of learning our MDKG representation. Inspired by the work of Lin *et al.* [20] which developed a GNN algorithm to learn representations of their KG. Therefore, we also adopted a GNN method to learn the representations of our MDKG to provide knowledgeable information for microbes and diseases.

For a microbe $m \in A$, we denoted $N^H(m)$ as the H-hop neighbors of m in the G, where H was a hyperparameter in our model. Due to the different numbers of the element in each hop neighborhood, we uniformly sampled a fixed size set $N_S^H(m)$ (sampled $N^H(m)$) to avoid complicated calculations. Note that, there existed duplicates in $N_S^H(m)$ when $|N^H(m)| < K$, where K represented the number of sampling neighbor size of each depth and was a hyperparameter in our method. We implemented the same operation and setting for a disease $d \in A$. Therefore, we can obtain the sampling neighbor set $N_S^H(d)$ of disease d in G. The representations of entities and relations in G were initially with random values which were generated via the Glorot uniform random generator [26]. $v_e(v_e \in R^{l'})$ and $v_r(v_r \in R^{l'})$ denoted the entity e which included microbes and diseases in G but not in A and the relation $r \in R$ representations, respectively, where l' was the dimension of representation and was a hyperparameter in our model. To distinguish the microbe (or the disease) in A and other entities in G, we defined v_m and v_d as the feature representations of microbe m and disease d in A. Then, we computed the importance C_r^m between a microbe feature representation v_m and a relation feature representation v_r that the relation r connected with microbe m in G as follows:

$$C_r^m = g(v_m * v_r), \tag{1}$$

where $*$ represented element-wise multiply, g denoted the non-linear activation function. Here, we adopted the Rectified Linear Unit(ReLU) as the activation function. To better capture the topological structure information of a microbe m, we calculated the neighborhood representation of microbe m as follows:

TABLE 3
The Number of Triplets Between Different Entity-Type Pairs in the MDKG

Entity-type pair	Database (resource)	Count
(Gene, Disease)	DRKG	95399
(Disease, Gene)	DRKG	27977
(Microbe, Metabolite)	VMH	818
	HMDB	235
(Metabolite, Disease)	VMH	115
	HMDB	27695
(Microbe, Microbe)	MIND	67780
(Compound, Disease)	DRKG	83895
(Disease, Disease)	DRKG	543
(Disease, Anatomy)	DRKG	3602
(Disease, Symptom)	DRKG	3357
(Disease, Organ)	VMH	559
(Microbe, Division)	NCBI	4856

$$v_{neighbor}^m = \sum_{e \in N_S^H(m)} C_{r_{m,e}}^m v_e, \tag{2}$$

where $N_S^H(m)$ demonstrated the sampled neighbor set of m in the H-hop neighbor set. Note that, $C_{r_{m,e}}^m$ denoted the microbe-relation importance, and $v_e(v_e \in R^{l'})$ was the representations of entity e. Therefore, we obtained the neighborhood representation of microbe m. For convenience, we implemented the same computation to gain the neighbor representation of disease d. After that, we aggregated the representation v_m of microbe m and its neighborhood representation $v_{neighbor}^m$. To simplify the calculation, we used the following aggregation operation to aggregate neighbors.

$$Agg_{v_m} = g((v_m || v_{neighbor}^m) \cdot W_{concat} + B_{concat}), \tag{3}$$

where $W_{concat} \in R^{2\,l' \times l'}$, $B_{concat} \in R^{l'}$ were learnable weight and bias parameters, respectively. Here, $\cdot$ represented the matrix multiply and $||$ was the concatenation operation. We adopted ReLU function as activation function g. Further, v_e was updated by aggregating the entity e's H-hop neighborhood information in G via the same aggregation operation. After that, we obtained the new representation $v_m = Agg_{v_m}$ of microbe m due to aggregating the updated v_e via the Equation (2). That could capture more topological information of microbe m by the information passing and aggregating in G. For convenience, the disease d new representation v_d was generated by the same operation with the microbe m.

3.3 Generating Representations of Microbes and Diseases

Based on the assumption that functionally similar microbes tended to show interaction or non-interaction characteristics with similar diseases and vice versa [7]. Therefore, we applied the Gaussian kernel similarity function on known microbe-disease associations to calculate Gaussian kernel similarity for diseases and microbes. For a certain microbe m, we defined the interaction profile which represented the interactions between the microbe m and all diseases in

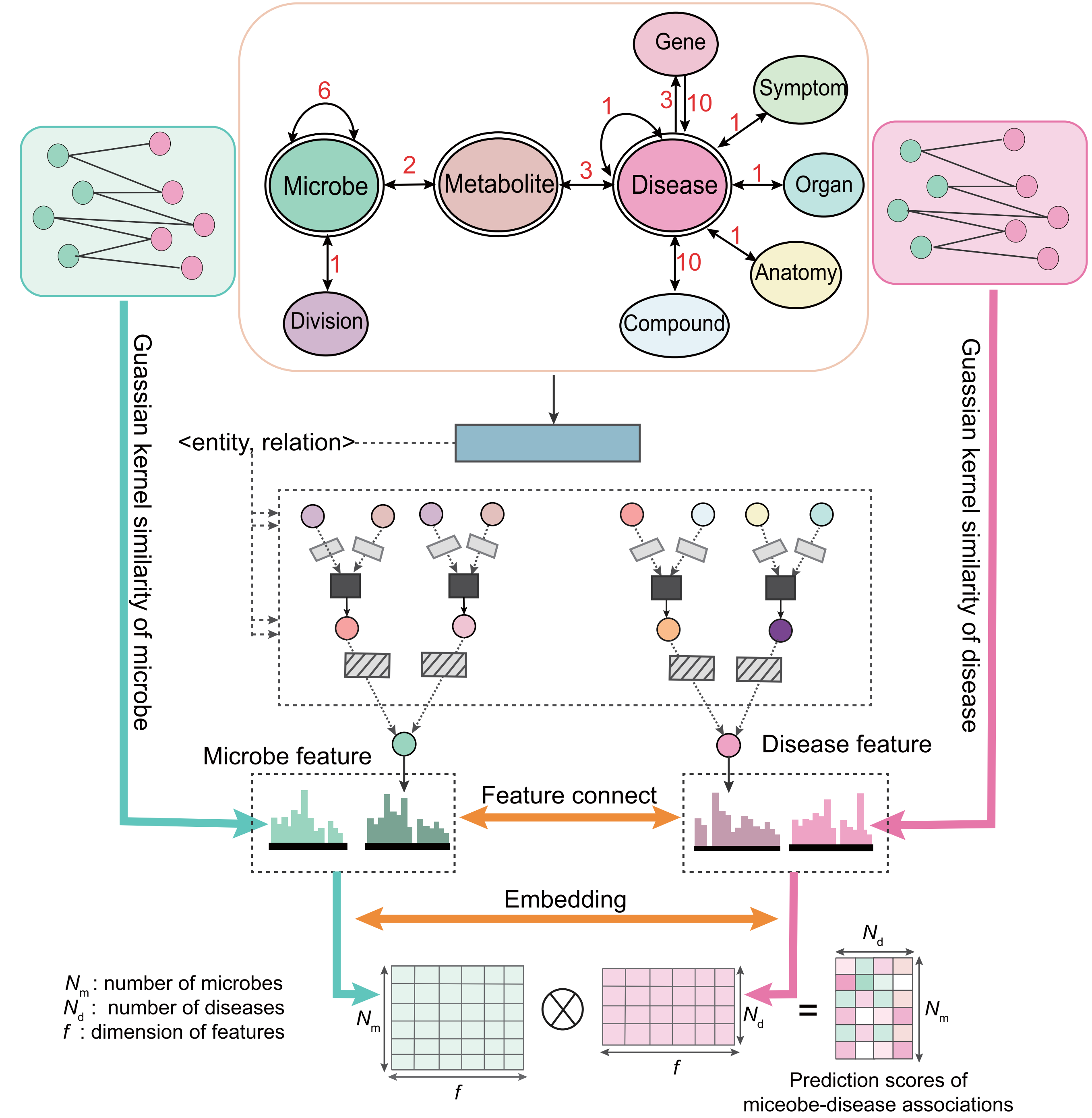


Fig. 1. The framework of our proposed KGNMDA model. First, we learned representations of microbes and diseases from MDKG which was in the middle of the top part using the GNN algorithm. Then, we combined Gaussian similarity features and representations learned from the previous step to generate the final representations of microbes and diseases. In the last part, we implemented a score function on final representations of microbes and diseases to predict scores of microbe-disease associations.

adjacent matrix A as $IP(m)$. Similarly, we defined the interaction profile which represented the interactions between the disease d and all microbes in adjacent matrix A as $IP(d)$. The Gaussian kernel similarities of diseases GD and microbes GM were calculated as follows:

$$GD(d,\hat{d}) = exp(-\lambda_{dd}||IP(d) - IP(\hat{d})||^2), \hat{d} \in D \setminus d \tag{4}$$

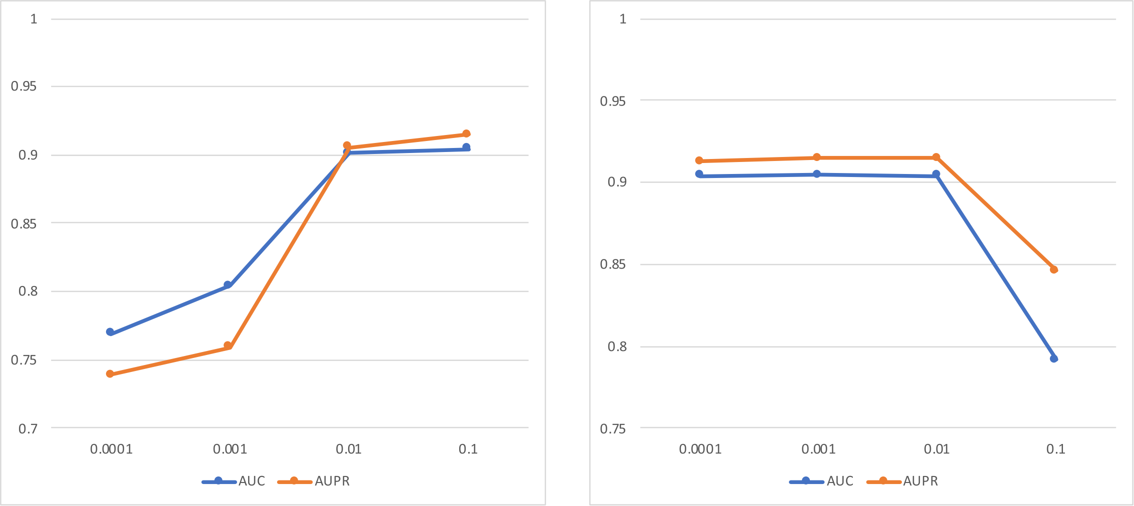
$$GM(m,\hat{m}) = exp(-\lambda_{mm}||IP(m) - IP(\hat{m})||^2), \hat{m} \in M \setminus m, \tag{5}$$

where M and D denoted the microbe set and disease set in A, respectively. $D \setminus d$ represented a set in D except d, and $M \setminus m$ denoted a set in M except m. Note λ_{dd} and λ_{mm} represented the normalized kernel bandwidths and were calculated as follows:

$$\lambda_{dd} = \lambda'_{dd/}\left(\frac{1}{nd}\sum_{d=1}^{nd}||IP(d)||^2\right), \tag{6}$$

$$\lambda_{mm} = \lambda'_{mm}/\left(\frac{1}{nm}\sum_{m=1}^{nm}||IP(m)||^2\right), \tag{7}$$

where λ'_{dd} and λ'_{mm} were the original bandwidths and were both set to 1. And nd and nm represented the number of



(a) L2 regularization coefficient

(b) Learning rate

Fig. 2. Parameter sensitivity under 5-fold CV on the HMDAD dataset (A) L2 regularization coefficient, (B) learning rate.

diseases and microbes, respectively. Then, we defined v'_m and v'_d as the final representations of microbe m and disease d, respectively. The process of computation was as follow:

$$v'_m = v_m || GM(m,*) \tag{8}$$

$$GM(m,*) = \underset{\hat{m} \in M \setminus m}{||} GM(m,\hat{m}) \tag{9}$$

$$v'_d = v_d || GD(d,*) \tag{10}$$

$$GD(d,*) = \underset{\hat{d} \in D \setminus d}{||} GM(d,\hat{d}), \tag{11}$$

where $||$ represented the concatenation operation. $GM(m,*)$ (or $GD(d,*)$) denoted the microbe m (or disease d) Gaussian similarity vector. Here, the dimension between v_m (or v_d) and $GM(m,*)$ (or $GD(d,*)$) is different, thus for convenience, we adopt the connection operation to combine that.

3.4 Microbe-Disease Associations Prediction

Finally, we proposed a score function to predict the scores of microbe-disease associations using representations of microbes and diseases

$$score(v'_d, v'_m) = (v'_d \cdot W_d) * (v'_m \cdot W_m) + B_{dm}, \tag{12}$$

where $W_d \in R^{l_d \times l''}$, $W_m \in R^{l_m \times l''}$ and $B_{dm} \in R^{l''}$ were learnable weight and bias parameters, respectively. l_d, l_m and l'' were the feature dimension of v'_d, v'_m and $score(v'_d, v'_m)$, separately. l'' was a hyperparameter in our model. In this model, we regarded microbe-disease associations prediction as a binary classification task. Therefore, we used binary cross-entropy as the loss function to optimize the model parameters

$$L = -y_{d,m} log \hat{y}_{d,m} - (1 - y_{d,m}) log(1 - \hat{y}_{d,m}), \tag{13}$$

where $\hat{y}_{d,m} = sigmoid(score(v'_d, v'_m))$ was the predicted score, $sigmoid(\cdot)$ represented the Sigmoid activation function. $y_{d,m}$ was the ground-truth value, and A represented the known microbe-disease associations. We adopted the Adam optimizer [27] for the optimization. Besides, we randomly sampled equal-size unknown disease-microbe pairs as negative samples for model training.

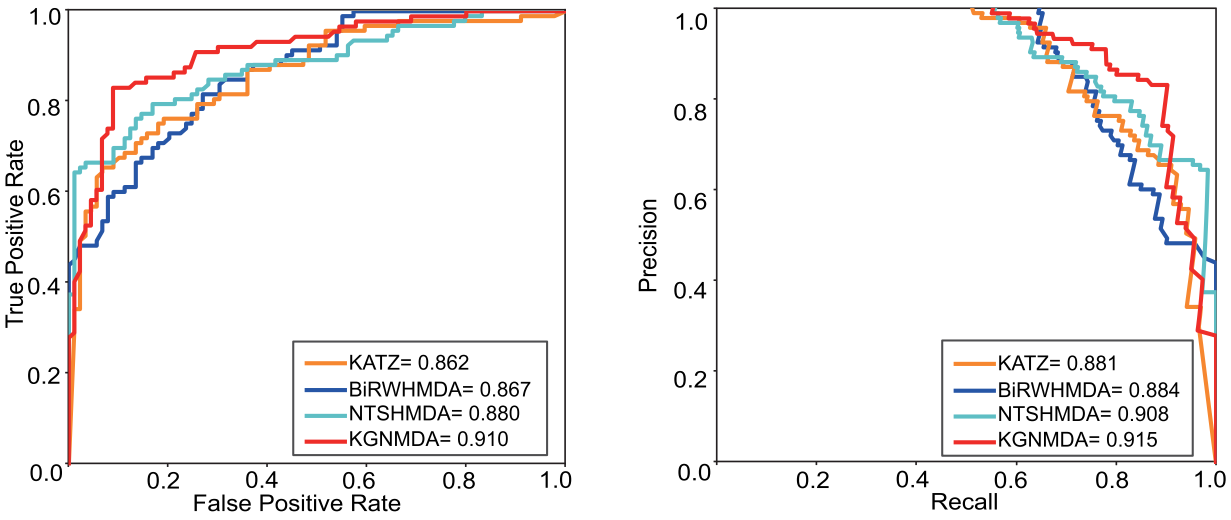
4 Results

4.1 Experiment Setup

We adopted the 5-fold cross-validation(CV) and 10-fold CV settings. For 5-fold CV and 10-fold CV, the known microbe-disease associations were randomly divided into five groups and ten groups, in each round test sample, included randomly sampled one group of microbe-disease pairs (i.e., positive samples) and an equal-size set of unknown microbe-disease associations (i.e., negative samples). To a positive or negative test sample (i.e., a microbe-disease pair), we considered that the model effectively predicted the microbe-disease pair if its score ranking was higher or lower than a given threshold. Hence, we set different thresholds to obtain the precision, true positive rates (TPR, sensitively/recall) and false positive rates (FPR, 1-specificity). Here, we calculated the percentage of the positive test samples in all samples predicted as positive test samples with the given threshold to obtain the precision. TPR represented the rate of the positive test samples whose score rankings were higher than the given threshold. Specificity was defined as the percentage of the negative test samples whose score rankings were ranked lower than the given threshold. Therefore, we could draw the receiver operating characteristics (ROC) curves by plotting TPR versus FPR at different thresholds. And, we could draw the precision-recall (PR) curves by plotting precision versus recall at different thresholds. Then, we adopted the area under the ROC curves (i.e., AUC) and the area under the PR curves (i.e., AUPR) as evaluation metrics to evaluate the performance of the model. To predict novel microbe–disease associations for new diseases and new microbes, respectively, we randomly sample 20% rows (5-fold-D) and 20% columns (5-fold-M) in the

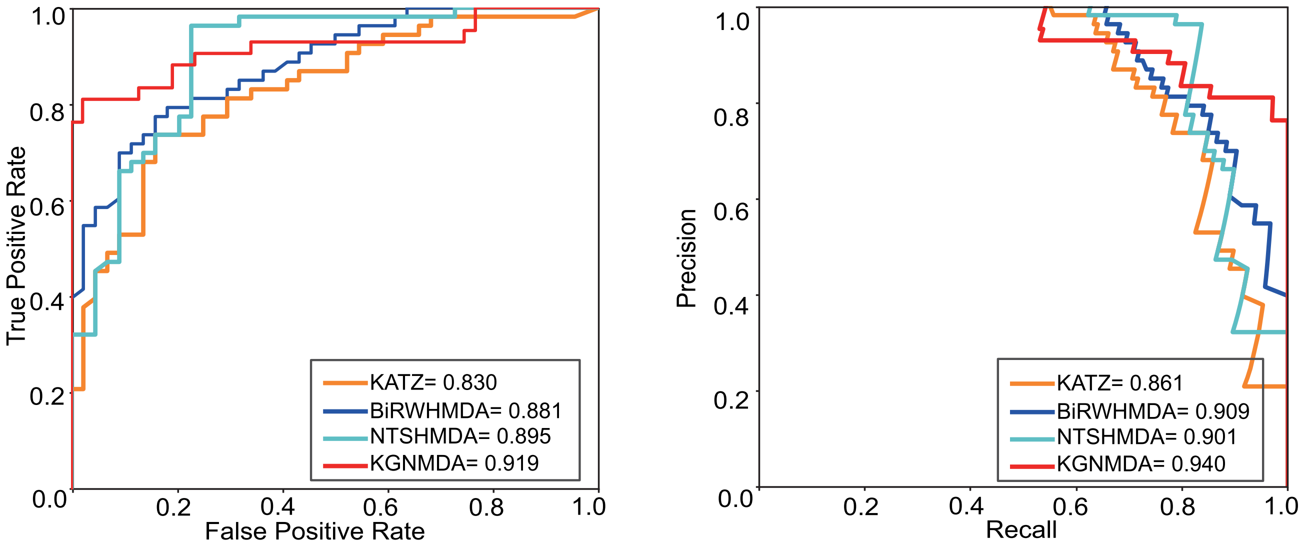
TABLE 4
The Details of the Model Parameters Setting

Parameter	Value	Parameter	Value
H	2	K	8
l'	32	l''	32
Epoch	50	Learning rate	1×10^{-3}
Batch size	32	L2 regularization coefficient	1×10^{-1}



(a) AUROC **(b) AUPR**

Fig. 3. The AUROC curve and AUPR curve of 5-fold CV on the HMDAD datasets between different methods.



(a) AUROC **(b) AUPR**

Fig. 4. The AUROC curve and AUPR curve of 10-fold CV on the HMDAD datasets between different methods.

adjacent matrix A as test samples, while the rest of the rows and columns are considered as training samples, respectively. We computed both metrics on the above cross-validation setups. To eliminate the influence of random division, we repeatedly implemented experiments 10 times to obtain the average prediction results.

4.2 Parameter Analysis

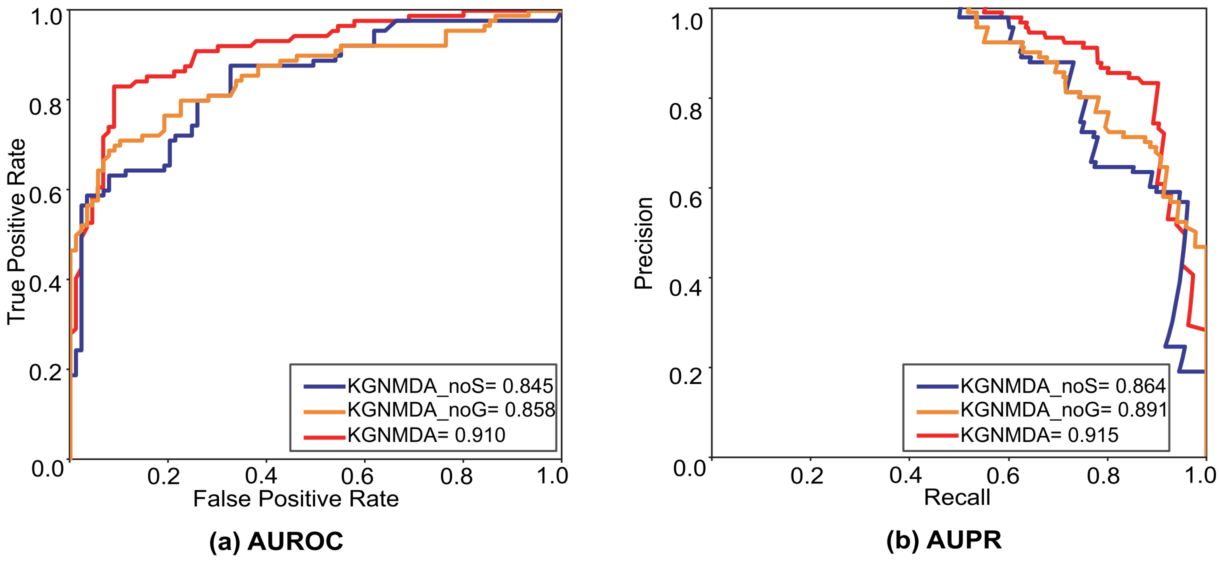
In our model, there are several parameters that influence the performance, such as the L2 regularization coefficient and learning rate. To evaluate the influences of these parameters, we implemented 5 fold-CV based on HMDAD dataset. We range L2 regularization coefficient and learning rate from 0.0001 to 0.1 with a step value of 10. As shown in Figs. 2 A and 2 B, we observed that the model gains the best performance when L2 regularization coefficient and learning rate were 0.1 and 0.001, respectively. Therefore, In our model, the details of model parameter settings were listed in Table 4.

4.3 Comparsion With Other Methods

We evaluated the performance of our proposed model via comparing with several methods that primarily contained matrix factorization-based method and network-based model. Including, KATAHMDA [7] was a network-based measurement method; NTSHMDA [10] proposed a random-walk based on network model; BiRWHMDA [9] developed a bi-random walk on the heterogeneous network method. Here, we conducted the above baseline methods on the HMDAD dataset on their default parameter settings. The

TABLE 5
The Results for 5-Fold-D and 5-Fold-M

Settings	AUC	AUPR
5-fold-D	0.6836	0.6232
5-fold-M	0.8685	0.8899



(a) AUROC

(b) AUPR

Fig. 5. The AUROC curve and AUPR curve of 5-fold CV on the HMDAD datasets between different variants of KGNMDA, i.e., KGNMDA_noG and KGNMDA_noS.

experiment results between baseline methods and our proposed model under in 5-fold CV and 10-fold CV on the HMDAD dataset were shown in Figs. 3 and 4. As shown in Figs. 3 and 4, we observed that our KGNMDA outperformed other models on AUC-ROC and AUC-PR experimental configurations. That demonstrated KGNMDA was effective in revealing the relationship between microbes and diseases. The reason for that may be the MDKG could provide more information for microbes and diseases and the GNN framework performed well on KG embedding. To evaluate KGNMDA on new diseases or new microbes that did not know any associations with microbes and diseases ahead, we implemented the 5-fold-D and 5-fold-M experiments. The results for 5-fold-D and 5-fold-M are shown in following table. As shown in the Table 5, we observed that KGNMDA could work for new microbes and new diseases. And the performance under 5-fold standard setting was significantly better than that under 5-fold-D and 5-fold-M settings. Because we had no known association pairs for new diseases and new microbes. Furthermore, since the count of microbes was bigger than the count of diseases that our model learned more abundant information from the microbes than the diseases, the performance of 5-fold-M was better than 5-fold-D.

4.4 Ablation Study

In this section, we introduced several variants of KGNMDA (i.e., KGNMDA_noG and KGNMDA_noS) to analyze the structure of the model. KGNMDA_noG represented that deleted the part of the Gaussian similarity features. KGNMDA_noS demonstrated that deleted the part of the score function, and directly calculated the score of microbe-disease associations via inputting the connection of microbes representation and diseases representation into the single-layer neural network. As shown in Fig. 5, we observed that KGNMDA outperformed KGNMDA_noG. That indicated the Gaussian similarity features help improve the performance. The result of KGNMDA_noS demonstrated that the score function in our model could capture more association information between microbes and diseases and help to improve the performance of prediction. Therefore, we considered that the Gaussian similarity features and score function both improved the model performance.

4.5 Case Studies

To further verify the prediction performance of our proposed model, we conducted case studies on two common diseases, i.e asthma and Inflammatory bowel disease(IBD). For each of them, we calculated the associated scores with all candidate microbes using our model and deleted the microbes which were known associated in the HMDAD dataset. Then we evaluated the performance by verifying the top-k (k=5, 10, 20) predicted microbes which were prioritized according to their scores from previous literature. Note that, if one microbe was associated with a given disease, the corresponding genus of the microbe was also assumed to be related to the disease.

To the disease asthma, it was a long-term inflammatory disease of the airways of the lungs and that was caused by genetic and environmental factors [28]. Accumulated literature showed that asthma was related to microorganisms in human bodies. For instance, the microbe Clostridium coccoides was associated with the disease asthma [29], which appeared in the prediction result of our proposed model. M Demirci *et al.* [5] found the microbe Faecalibacterium prausnitzii levels were reduced in the gut microbiota of children with allergic asthma. As shown in Table 6, we can observe that 5, 8, and 16 disease-microbe associations were verified from previous literature among the top 5, 10, and 20 predicted candidate microbes of disease asthma. In addition, IBD was a broad term that describes conditions characterized by chronic inflammation of the gastrointestinal tract, which included Crohn disease and ulcerative colitis [30]. Accumulated researches demonstrated various microbes were related to IBD. For example, the microbe Prevotella was found that was related to the disease IBD [31]. Harry *et al.* [32] found that microbes Bacteroidetes and Faecalibacterium prausnitzii were shown low counts in the IBD's patients. In Table 7, there were 5, 10, and 18 disease-microbe associations were confirmed from previous researches among the top 5, 10, and 20 predicted IBD-related microbes.

TABLE 6
The Details of Prediction Results of the Top 20 Asthma-Associated Microbes

Rank	Microbes	Evidence
1	Firmicutes	PMID: 23265859
2	Actinobacteria	PMID: 23265859
3	Lachnospiraceae	PMID: 31958431
4	Clostridium coccoides	PMID: 21477358
5	Staphylococcus aureus	PMID: 21477358
6	Clostridium difficile	PMID: 21872915
7	Enterobacteriaceae	PMID: 28947029
8	Bacteroides	PMID: 18822123
9	Bacteroides vulgatus	Unconfirmed
10	Bacteroides uniformis	Unconfirmed
11	Faecalibacterium prausnitzii	PMID: 30765132
12	Clostridium	PMID: 21477358
13	Lactobacillus	PMID: 30400588
14	Bacteroidaceae	PMID: 28947029
15	Burkholderia	PMID: 24451910
16	Stenotrophomonas maltophilia	Unconfirmed
17	Escherichia coli	Unconfirmed
18	Enterococcus	PMID: 29788027
19	Clostridium leptum	PMID: 26565810
20	Clostridia	PMID: 21477358

TABLE 7
The Details of Prediction Results of the Top 20 Inflammatory Bowel disease(IBD)-Associated Microbes

Rank	Microbes	Evidence
1	Prevotella	PMID: 25307765
2	Firmicutes	PMID: 25307765
3	Bacteroidetes	PMID: 25307765
4	Clostridium coccoides	PMID: 19235886
5	Staphylococcus aureus	PMID: 11424320
6	Clostridium difficile	PMID: 27499718
7	Haemophilus	PMID: 24013298
8	Enterobacteriaceae	PMID: 30172257
9	Bacteroides	PMID: 25307765
10	Helicobacter pylori	PMID: 22221289
11	Bacteroides vulgatus	PMID: 21575910
12	Bacteroides uniformis	PMID: 28405140
13	Faecalibacterium prausnitzii	PMID: 19235886
14	Clostridium	Unconfirmed
15	Lactobacillus	PMID: 26340825
16	Bacteroidaceae	PMID: 17897884
17	Burkholderia	PMID: 20237101
18	Stenotrophomonas maltophilia	Unconfirmed
19	Escherichia coli	PMID: 17440180
20	Staphylococcus	PMID: 11424320

That further demonstrated that our model showed great prediction performance and was useful in searching disease-associations microbes.

5 Conclusion

Recovering the microbe-disease relations can assist to understand the complex pathogenic mechanism of human diseases, and provide new insight for microbe-oriented medicine. For example. K. Shahanavaj *et al.* [37] demonstrated microbiota can be biomarkers for cancer and potentially help to disclose a novel paradigm of research for the treatment of cancer. Diseases can be treated by intervention in the human microbiome through medications or other living conditions (such as diet) [38]. The conventional wet lab methods of identifying microbe-disease associations were time-consuming, expensive, and labor-intensive. Therefore, computational methods were important for helping recover the microbe-disease relations. However, previous computational models did not systematically integrate the biological knowledge information of microorganisms and diseases. That may be not conducive to the prediction of the associations between new microorganisms and new diseases in the future.

In this study, we proposed a novel model named KGNMDA for human microbe-disease association prediction. First, we constructed an MDKG from existed several databases to provide biomedical information for microbes and diseases, The MDKG contained metabolites, genes, organs, and some other biomedical entities which were related to microbes and diseases. Note that, we were the first attempt to adopt KG to help to predict microbe-disease associations. Second, we adopted the GNN algorithm to learn representations of our MDKG. Then we combined the Gaussian kernel interaction profile similarity and learned representations from MDKG to generate the final representations of microbes and diseases. Finally, we developed a score function to predict scores of microbe-disease associations based on final representations of microbes and diseases. Comprehensive experiments demonstrated that our model showed good prediction performance and was useful in searching disease-associations microbes.

However, there were still some limitations in our model. On the one hand, our MDKG was still incomplete. Therefore, in the future, we will collect or mine more biological entities that were related to microbes or diseases to extend MDKG. On the other hand, our model can not predict the type of associations between different microbe-disease pairs. Hence, we consider adding the information of relationship types into our model to achieve that in our future work.

Acknowledgments

Changzhi Jiang and Minli Tang are contributed equally to this work.

References

[1] Z. Wen, C. Yan, G. Duan, S. Li, F.-X. Wu, and J. Wang, "A survey on predicting microbe-disease associations: Biological data and computational methods," *Brief. Bioinf.*, vol. 22, pp. 1–20, 2020.

[2] E. Holmes, A. Wijeyesekera, S. D. Taylor-Robinson, and J. K. Nicholson, "The promise of metabolic phenotyping in gastroenterology and hepatology," *Nature Rev. Gastroenterol. Hepatol.*, vol. 12, no. 8, pp. 458–471, 2015.

[3] S. V. Lynch and O. Pedersen, "The human intestinal microbiome in health and disease," *New England J. Med.*, vol. 375, no. 24, pp. 2369–2379, 2016.

[4] M. I. El Mouzan *et al.*, "Microbiota profile in new-onset pediatric crohn's disease: Data from a non-western population," *Gut Pathogens*, vol. 10, no. 1, pp. 1–10, 2018.

[5] M. Demirci *et al.*, "Reduced akkermansia muciniphila and faecalibacterium prausnitzii levels in the gut microbiota of children with allergic asthma," *Allergologia et immunopathologia*, vol. 47, no. 4, pp. 365–371, 2019.

[6] R. F. Schwabe and C. Jobin, "The microbiome and cancer," *Nature Rev. Cancer*, vol. 13, no. 11, pp. 800–812, 2013.

[7] X. Chen, Y.-A. Huang, Z.-H. You, G.-Y. Yan, and X.-S. Wang, "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, 2017.

[8] Y. Long and J. Luo, "WMGHMDA: A novel weighted meta-graph-based model for predicting human microbe-disease association on heterogeneous information network," *BMC Bioinf.*, vol. 20, no. 1, pp. 1–18, 2019.

[9] S. Zou, J. Zhang, and Z. Zhang, "A novel approach for predicting microbe-disease associations by bi-random walk on the heterogeneous network," *PLoS One*, vol. 12, no. 9, pp. 1–16, 2017.

[10] J. Luo and Y. 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.

[11] Z. Shen, Z. Jiang, and W. Bao, "CMFHMDA: Collaborative matrix factorization for human microbe-disease association prediction," in *Proc. Int. Conf. Intell. Comput.*, 2017, pp. 261–269.

[12] B.-S. He, L.-H. Peng, and Z. Li, "Human microbe-disease association prediction with graph regularized non-negative matrix factorization," *Front. Microbiol.*, vol. 9, pp. 1–11, 2018.

[13] G. Duan, C. Yan, F. 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.

[14] Y. Long, J. Luo, Y. Zhang, and Y. Xia, "Predicting human microbe–disease associations via graph attention networks with inductive matrix completion," *Brief. Bioinf.*, vol. 22, pp. 1–13, 2020.

[15] S. Ji, S. Pan, E. Cambria, P. Marttinen, and P. S. Yu, "A survey on knowledge graphs: Representation, acquisition and applications," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 2, pp. 494–514, 2021.

[16] T. Lee *et al.*, "Oral versus intravenous iron replacement therapy distinctly alters the gut microbiota and metabolome in patients with IBD," *Gut*, vol. 66, no. 5, pp. 863–871, 2017.

[17] S. K. Mohamed, V. Nováček, and A. Nounu, "Discovering protein drug targets using knowledge graph embeddings," *Bioinformatics*, vol. 36, no. 2, pp. 603–610, 2020.

[18] A. Bordes, N. Usunier, A. Garcia-Duran, J. Weston, and O. Yakhnenko, "Translating embeddings for modeling multi-relational data," in *Proc. Adv. Neural Inf. Process. Syst.*, 2013, pp. 2787–2795.

[19] J. Zhou *et al.*, "Graph neural networks: A review of methods and applications," *AI Open*, vol. 1, pp. 57–81, 2020.

[20] X. Lin, Z. Quan, Z.-J. Wang, T. Ma, and X. Zeng, "KGNN: Knowledge graph neural network for drug-drug interaction prediction," in *Proc. 29th Int. Joint Conf. Artif. Intell.*, pp. 2739–2745, 2020.

[21] W. Ma *et al.*, "An analysis of human microbe–disease associations," *Brief. Bioinf.*, vol. 18, no. 1, pp. 85–97, 2017.

[22] V. N. Ioannidis *et al.*, "DRKG - Drug repurposing knowledge graph for Covid-19," 2020. [Online]. Available: https://github.com/gnn4dr/DRKG/

[23] A. Noronha *et al.*, "The Virtual Metabolic Human database: Integrating human and gut microbiome metabolism with nutrition and disease," *Nucleic Acids Res.*, vol. 47, no. D1, pp. D614–D624, Oct. 2018. [Online]. Available: https://doi.org/10.1093/nar/gky992

[24] D. S. Wishart *et al.*, "HMDB 4.0: The human metabolome database for 2018," *Nucleic Acids Res.*, vol. 46, no. D1, pp. D608–D617, 2018.

[25] D. L. Wheeler *et al.*, "Database resources of the national center for biotechnology information," *Nucleic Acids Res.*, vol. 36, no. suppl_1, pp. D13–D21, 2007.

[26] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proc. 13th Int. Conf. Artif. Intell. Statist.*, 2010, pp. 249–256.

[27] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.

[28] F. Martinez, "Genes, environments, development and asthma: A reappraisal," *Eur. Respir. J.*, vol. 29, no. 1, pp. 179–184, 2007.

[29] C. Vael, L. Vanheirstraeten, K. N. Desager, and H. Goossens, "Denaturing gradient gel electrophoresis of neonatal intestinal microbiota in relation to the development of asthma," *BMC Microbiol.*, vol. 11, no. 1, pp. 1–7, 2011.

[30] N. J. Talley and S. O'Connor, *Clinical Examination: A Systematic Guide to Physical Diagnosis*. Chatswood, NSW, Australia: Elsevier, 2022.

[31] W. A. Walters, Z. Xu, and R. Knight, "Meta-analyses of human gut microbes associated with obesity and IBD," *FEBS Lett.*, vol. 588, no. 22, pp. 4223–4233, 2014.

[32] H. Sokol *et al.*, "Low counts of faecalibacterium prausnitzii in colitis microbiota," *Inflammatory Bowel Dis.*, vol. 15, no. 8, pp. 1183–1189, 2009.

[33] W. Barcik, R. C. Boutin, M. Sokolowska, and B. B. Finlay, "The role of lung and gut microbiota in the pathology of asthma," *Immunity*, vol. 52, no. 2, pp. 241–255, 2020.

[34] D. W. Bang *et al.*, "Asthma and risk of non-respiratory tract infection: A population-based case–control study," *BMJ Open*, vol. 3, no. 10, pp. 1–8, 2013.

[35] Y. Zhou *et al.*, "The upper-airway microbiota and loss of asthma control among asthmatic children," *Nature Commun.*, vol. 10, no. 1, pp. 1–10, 2019.

[36] E. Buendía, J. Zakzuk, H. San-Juan-Vergara, E. Zurek, N. J. Ajami, and L. Caraballo, "Gut microbiota components are associated with fixed airway obstruction in asthmatic patients living in the tropics," *Sci. Rep.*, vol. 8, no. 1, pp. 1–12, 2018.

[37] K. Shahanavaj *et al.*, "Cancer and the microbiome: Potential applications as new tumor biomarker," *Expert Rev. Anticancer Ther.*, vol. 15, no. 3, pp. 317–330, 2015.

[38] J. R. Lukens *et al.*, "Dietary modulation of the microbiome affects autoinflammatory disease," *Nature*, vol. 516, no. 7530, pp. 246–249, 2014.

Changzhi Jiang is currently working toward the PhD degree with the Department of Computer Science and Technology, School of Informatics, Xiamen University. His main research interests include bioinformatics and drug discovery.

Minli Tang is currently working toward the PhD degree with the Department of Computer Science and Technology, School of Informatics, Xiamen University. Her main research interests include deep learning and computer vision.

Shuting Jin is currently working toward the PhD degree with the Department of Computer Science and Technology, School of Informatics, Xiamen University. Her main research interests include bioinformatics and drug discovery.

Wei Huang is currently working toward a master's degree with the Department of Computer Science and Technology, School of Informatics, Xiamen University. His main research interests include bioinformatics and drug discovery.

Xiangrong Liu is currently a professor with the Department of Computer Science and Technology, School of Informatics, Xiamen University. His main research interests include computational intelligence, computational theory, data mining, biological information processing, mobile and micro-sensing technology, etc.