MADGAN:A microbe-disease association prediction model based on generative adversarial networks

Weixin Hu[1], Xiaoyu Yang[2], Lei Wang[2,3]* and Xianyou Zhu[1]*

[1]College of Computer Science and Technology, Hengyang Normal University, Hengyang, China, [2]Institute of Bioinformatics Complex Network Big Data, Changsha University, Changsha, China, [3]Big Data Innovation and Entrepreneurship Education Center of Hunan Province, Changsha University, Changsha, China

Researches have demonstrated that microorganisms are indispensable for the nutrition transportation, growth and development of human bodies, and disorder and imbalance of microbiota may lead to the occurrence of diseases. Therefore, it is crucial to study relationships between microbes and diseases. In this manuscript, we proposed a novel prediction model named MADGAN to infer potential microbe-disease associations by combining biological information of microbes and diseases with the generative adversarial networks. To our knowledge, it is the first attempt to use the generative adversarial network to complete this important task. In MADGAN, we firstly constructed different features for microbes and diseases based on multiple similarity metrics. And then, we further adopted graph convolution neural network (GCN) to derive different features for microbes and diseases automatically. Finally, we trained MADGAN to identify latent microbe-disease associations by games between the generation network and the decision network. Especially, in order to prevent over-smoothing during the model training process, we introduced the cross-level weight distribution structure to enhance the depth of the network based on the idea of residual network. Moreover, in order to validate the performance of MADGAN, we conducted comprehensive experiments and case studies based on databases of HMDAD and Disbiome respectively, and experimental results demonstrated that MADGAN not only achieved satisfactory prediction performances, but also outperformed existing state-of-the-art prediction models.

1. Introduction

Microbes are far more numerous than human cells (Integrative HMP (iHMP) Research Network Consortium, 2014; Sender et al., 2016), and play an important role in human beings (Human Microbiome Project Consortium, 2012). The microorganisms parasitic on the human body constitute the human microbial community, and their composition varies from person to person (Human Microbiome Project Consortium, 2012). These microbial populations can not only protect the human body from foreign microorganisms and pathogens, but also participate in intestinal digestion and absorption, and promote metabolism (Guarner and Malagelada, 2003;

Kau et al., 2011). Therefore, to some extent, the human microbial population can even be regarded as human "forgotten organs"(Quigley, 2013), the imbalance of microorganisms will not only lead to the occurrence of nervous system diseases, but also affect the immune and metabolic functions of the human body (Cenit et al., 2017; Li et al., 2017). For example, changes in intestinal microbiota are highly correlated with the pathogenesis of various nervous system diseases, including depression, autism (Kim et al., 2018), asthma (Al-Moamary et al., 2021) and cancer (Schwabe and Jobin, 2013), etc. Of course, there is also evidence showing that microbial populations can help regulate disease as well (Cryan and Dinan, 2012). For instance, researches show that lactic acid bacteria and bifid bacteria play a positive role in regulating anxiety, cognition, pain and depression symptoms (Desbonnet et al., 2010). In addition, Huang pointed out that microorganisms can affect the hypersensitivity and asthma of susceptible people. Early intervention to promote the healthy composition of human microbiome may help prevent asthma (Huang, 2013). Hence, it is meaningful to infer potential relationships between microorganisms and diseases, which can not only help researchers understand the pathogenesis of diseases, but also help us to prevent, diagnose and treat diseases, thus promoting global human health. Utilizing biotechnology to identify microbe-disease associations is time-consuming, costly and blind, so it is meaningful to identify potential microbe-disease associations through computational methods. Up to now, representative calculative methods can be roughly divided into four categories, such as the network-based, binary local features-based, matrix factorization/completion-based and graph neural network-based methods. Among them, the network-based methods infer latent microbe-disease associations by mainly adopting the topology information of different networks. For example, Chen et al. (2017) proposed a KATZ-based model KATZHMDA to infer possible microbe-disease associations based on a newly constructed heterogeneous network, which scores potential disease related microbes by step size and path numbers. Zeng et al. (2022) introduced the knowledge graph into the field of drug discovery, integrated data information through a displayed structure, and strengthened the structured connection and semantic relationship between entities. However, the methods based on binary local features focus on taking microbes and diseases as local objects, and identify potential microbe-disease associations by combining the features between them. For instance, Huang et al. (2017) developed a combined recommendation algorithm based on neighborhood and graph by integrating two independent recommendation models to recommend disease related microbes. In addition, Matrix factorization/completion-based methods aim to decompose the known incidence matrix into two characteristic matrices, and approximate the incidence matrix with the product of the two matrices. For instance, Shen et al. (2017) proposed a matrix factorization-based model for microbe-disease association prediction, which integrated known microbe-disease associations and introduced a collaborative matrix factorization scheme to update the correlation matrix about microbes and diseases for inferring the most possible disease-related microbes. Finally, the graph neural network-based methods used to learn structural data by taking microbe and disease related data as the input of the neural networks, so as to extract and explore features and patterns in graph structural data. For example, Long et al. (2021) developed a graph attention network with inductive matrix completion to detect potential microbe-disease associations. Cheng

et al. (2021) used the deep generative model as an entry point to discuss and study the *de novo* molecular design for drug discovery (*de novo* molecular design for drug discovery).

The emergence of generative adversarial networks is another milestone in the field of computer vision. It provides a new tool for solving various image prediction problems. For instance, in 2014, Lan et al. proposed a framework for estimating the generative adversarial network model through the confrontation process, and improved the ability of the model through the mutual game between generative adversarial networks (Goodfellow et al., 2020). However, the generative adversarial network still has problems such as unstable results and difficult training. Hence, Arjovsky et al. (2017) conducted a theoretical analysis of the generative adversarial network and provided an optimal solution. Later, new results appeared in the field of image processing, such as Style GAN (Karras et al., 2019), Cycle GAN (Zhu et al., 2017), SeCGAN (Wu et al., 2019), etc. In recent years, many researchers have begun to explore the application of generative adversarial networks in other fields. For example, Lei et al. (2019) applied it in the direction of dynamic information generation to build a nonlinear time link prediction model. Dai et al. (2021) introduced generative adversarial networks to natural language translation work. Zheng et al. (2022) utilized a generative adversarial network model to predict urban traffic flow.

In this paper, a generative adversarial network framework called MADGAN was designed for latent microbe-disease association prediction, in which, a GCN was adopted to obtain the microbe-disease association features first, and then, we would train the ability of MADGAN by games between the generation network and the decision network. And at the same time, inspired by the idea of residual network, we introduced the cross-level weight distribution structure to enhance the depth of the network to prevent over-smoothing during the model training process. Finally, intensive experiments based on the *k*-fold cross-validation framework were implemented to compare the prediction performance between MADGAN and state-of-the-art prediction models. And as a result, MADGAN was proved to be of satisfactory prediction ability and outperformed existing representative competing models.

2. Materials and methods

2.1. Construction of the microbe-disease association network

In this section, we would download known microbe-disease associations from two well-known public databases including HMDAD (Ma et al., 2017) and Disbiome (Janssens et al., 2018) respectively. Among them, HMDAD[1] is the first microbe-disease association database constructed by ma et al. in 2017, which contains 483 known microbe-disease associations. After removing duplicate data, we finally obtained 450 different known microbe-disease associations between 39 diseases and 292 microbes. Besides, Disbiome[2] is a public microbe-disease association database

---

1 http://www.cuilab.cn/hmdad

2 https://disbiome.ugent.be/home

constructed by Janssens et al., in which, there are 5,573 known associations between 240 diseases and 1,098 microbes collected from published academic papers. After removing duplicate data, we finally derived 4,351 known microbe-disease associations between 218 diseases and 1,052 microbes. For convenience, let n_d and n_m denote the numbers of newly-downloaded diseases and microbes respectively, then we can obtain a adjacency matrix $A \in \mathbb{R}^{n_d \times n_m}$ as follows: for any given disease d_i and a microbe m_j, if there is a known association between them, there is $A_{ij} = 1$, otherwise, there is $A_{ij} = 0$.

2.2. Multiple similarity calculation of disease

2.2.1. Gaussian interaction profile kernel similarity of disease

Based on the assumption that two similar diseases will show similar interaction and non-interaction relationship with the same microorganism (Chen et al., 2017), in this section, we will first calculate the Gaussian interaction profile kernel similarity between a pair of diseases d_i and d_j as follows:

$$GD(d_i, d_j) = \exp\left(-\lambda_d \parallel A(i,:) - A(j,:) \parallel^2\right) \qquad (1)$$

Where $A(i,:)$ and $A(j,:)$ represent the ith and j^{th} rows of the adjacency matrix A respectively, and λ_d denotes the normalized kernel bandwidths that can be calculated as follows:

$$\lambda_d = \frac{1}{\left(\frac{1}{n_d}\sum_{i=1}^{n_d} \parallel A(i,:) \parallel^2\right)} \qquad (2)$$

2.2.2. Cosine similarity of disease

Based on the assumption that if two diseases are similar to each other, then their cosine curves will be more coincident, in this section, we will define the cosine similarity between a pair of diseases d_i and d_j as follows:

$$CD(d_i, d_j) = \left(A(i,:) \cdot A(j,:)\right) / \left(\mid A(i,:) \mid * \mid A(j,:) \mid\right) \qquad (3)$$

The result of cosine similarity has good stability and certainty, the calculation speed is fast and the result is more intuitive. Suitable for large-scale information retrieval. Where $A(i,:) \cdot A(j,:)$ denotes multiplying the vectors of row i and row j, $\mid A(i,:) \mid$ represents the mode of $A(i,:)$, and $\mid A(j,:) \mid$ represents the mode of $A(j,:)$. $\mid A(i,:) \mid * \mid A(j,:) \mid$ represents the multiplication of two moduli, and then the value of the modulus is removed by the product of the vector, and finally the cosine value of the angle between the two diseases is obtained, that is, the cosine similarity. The calculation result of cosine similarity is between −1 and 1. When the similarity between two diseases is extremely high, the calculation result tends to be 1. When the similarity between two diseases is very low, the calculation result tends to −1.

2.2.3. Functional similarity of disease

Based on the assumption that similar diseases tend to interact with similar genes, in this section, we will calculate the disease functional similarity based on the functional associations between disease-related genes (Xu and Li, 2006; Wei and Liu, 2020) as follows: Firstly, we download the gene interactions from HumanNet database[3], in which, every interaction has an associated log-likelihood score (LLS). And then, for any given diseases d_i and d_j, let $G_i = \{g_{i_1}, g_{i_2}, \ldots, g_{i_m}\}$ and $G_j = \{g_{j_1}, g_{j_2}, \ldots, g_{j_n}\}$ denote the newly-obtained gene sets of d_i and d_j separately, we will define the functional similarity between d_i and d_j as follows:

$$DFS(d_i, d_j) = \frac{\sum_{g_k \in G_i} F_{G_j}(g_k) + \sum_{g_k \in G_j} F_{G_i}(g_k)}{m + n} \qquad (4)$$

Where $F_{G_i}(g_p) = \max_{g_q \in G_i}\left(FSS(g_p, g_q)\right)$, and $FSS(g_p, g_q)$ is the functional similarity score between the genes g_p and g_q, which can be calculated as follows:

$$FSS(g_p, g_q) = \begin{cases} 1 & if \ p = q \\ \dfrac{LLS(g_p, g_q) - LLS_{\min}}{LLS_{\max} - LLS_{\min}} & if \ p \neq q \end{cases} \qquad (5)$$

Where $LLS_{\max}$ and $LLS_{\min}$ represent the maximum value of LLS and the minimum value of LLS in HumanNet, respectively.

Thereafter, by combining above GIP kernel similarity, disease cosine similarity and functional similarity of disease, we can obtain an integrated similarity matrix of disease as follows:

$$DS = \frac{GD + CD + DFS}{3} \qquad (6)$$

2.3. Multiple similarity calculation of microbe

2.3.1. Gaussian interaction profile kernel similarity of microbe

In the same way, we can calculate the gaussian interaction profile kernel similarity between any two microbes m_i and m_j as follows:

$$MD(m_i, m_j) = \exp\left(-\lambda_m \parallel A(:,i) - A(:,j) \parallel^2\right) \qquad (7)$$

Where $A(:,i)$ and $A(:,j)$ represent the ith and j^{th} columns of the adjacency matrix A respectively, and λ_m denotes the normalized kernel bandwidths that can be calculated as follows:

$$\lambda_m = \frac{1}{\left(\frac{1}{n_m}\sum_{i=1}^{n_m} \parallel A(:,i) \parallel^2\right)} \qquad (8)$$

---

3 https://www.inetbio.org/humannet

2.3.2. Cosine similarity of microbe

Similarly, the cosine similarity between any two microbes m_i and m_j can be obtained as follows:

$$CM\left(m_i,m_j\right)=\left(A(:,i)\cdot A(:,j)\right)/\left(|A(:,i)|\times|A(:,j)|\right) \quad (9)$$

The calculation process of cosine similarity between two microorganisms is the same as that of disease cosine similarity. Similarly, when the similarity between two microorganisms is extremely high, the calculation result tends to be 1. When the similarity between two microorganisms is very low, the calculation result tends to −1.

2.3.3. Functional similarity of microbe

In this section, we will calculate the functional similarity of microbe by using the following method proposed in the reference (Zhang et al., 2018): for any given disease d_t, it is first represented by a Directed Acyclic Graph $DAG_{d_t}=\left(V_{d_t},E_{d_t}\right)$, where V_{d_t} includes the disease d_t and its ancestor diseases, E_{d_t} contains all the directed edges from parent nodes to children nodes (Wang et al., 2010), and then, the semantic contribution of the disease d_l in V_{d_t} to d_t is defined as:

$$SC_{d_t}\left(d_i\right)=$$
$$\begin{cases} 1 & if\ d_l=d_t \\ \max\left\{0.5\times SC_{d_t}\left(d_l'\right)|d_l'\in children\ of\ d_l\right\} & otherwise \end{cases} \quad (10)$$

The semantic value of disease d_t is formulated by:

$$SV_{d_t}=\sum_{d_i\in V_{d_t}}SC_{d_t}\left(d_l\right) \quad (11)$$

Then, the semantic similarity between any two diseases d_i and d_j can be defined as follows:

$$DSS\left(d_i,d_j\right)=\frac{\sum_{d_i\in V_{d_i}\cap V_{d_j}}\left(SC_{d_i}\left(d_l\right)+SC_{d_j}\left(d_l\right)\right)}{SV_{d_i}+SV_{d_j}} \quad (12)$$

Besides, based on above formulae, we can further define the similarity between the disease d_i and a set of diseases D as follows:

$$DS\left(d_i,D\right)=\max_{d_j\in D}\left(DSS\left(d_i,d_j\right)\right) \quad (13)$$

Hence, for any two given microbes m_i and m_j, we can calculate the function similarity between them as follows:

$$MFS\left(m_i,m_j\right)=\frac{\sum_{d_j\in D_j}DS\left(d_j,D_i\right)+\sum_{d_j\in D_i}DS\left(d_j,D_j\right)}{|D_i|+|D_j|} \quad (14)$$

Where D_i denotes the set of diseases associated with the microbe m_i, and D_j represents the set of diseases associated with the microbe m_j.

Obviously, by combining above GIP kernel similarity, disease cosine similarity and functional similarity of microbe, we can obtain an integrated similarity matrix of microbe as follows:

$$MS=\frac{MD+CM+MFS}{3} \quad (15)$$

2.4. Construction of the heterogeneous network

Based on above descriptions, it is easy to see that we can construct a heterogeneous network Y through integrating the integrated similarity matrix DS of disease and the integrated similarity matrix MS of microbe with the adjacency matrix A as follows:

$$Y=\begin{bmatrix} DS & A \\ A^T & MS \end{bmatrix} \quad (16)$$

3. Methods

The main framework of this paper is generative adversarial networks. A generative adversarial network consists of a generative network and a decision network, and it works by enhancing the model's capabilities during the mutual gaming of the two networks. As shown in Figure 1, the information of known microbial-disease association data is extracted from the database, and after the calculation of similarity, it is input into the generative network. The core of the generative network consists of a GCN layer and an attention mechanism, which consists of a graph convolutional layer and a sparse graph convolutional layer. The data are passed through the generative network to generate prediction results, and the prediction results and the original sample data are input into the discriminator, which distinguishes the real results from the generated results and returns to update the model parameters of the generative network. This is a game process, in which the generative network needs to generate prediction results that are sufficient to confuse the judgment of the discriminator, while the discriminator needs to correctly distinguish the generated results from the true results. The ability of the generative network model is continuously improved during the game until the discriminator and the generative network reach an equilibrium, i.e., the probability of both the predicted and true outcomes is one half.

The generator network uses the information of the data set to output data samples, and the generator $G(\bullet)$ obtains a random sample z from the data samples, and z conforms to the $p(z)$ probability distribution. After the generator generates data, it will be sent to the discriminator $D(\bullet)$, and the discriminator will try to predict the authenticity of the data after receiving real data or generated data. At the same time, it also needs a sample x from the real data distribution $p_{data}(x)$, the discriminator uses the activation function to solve a binary classification task, and outputs a value of 0–1 to distinguish the real result from the predicted result.
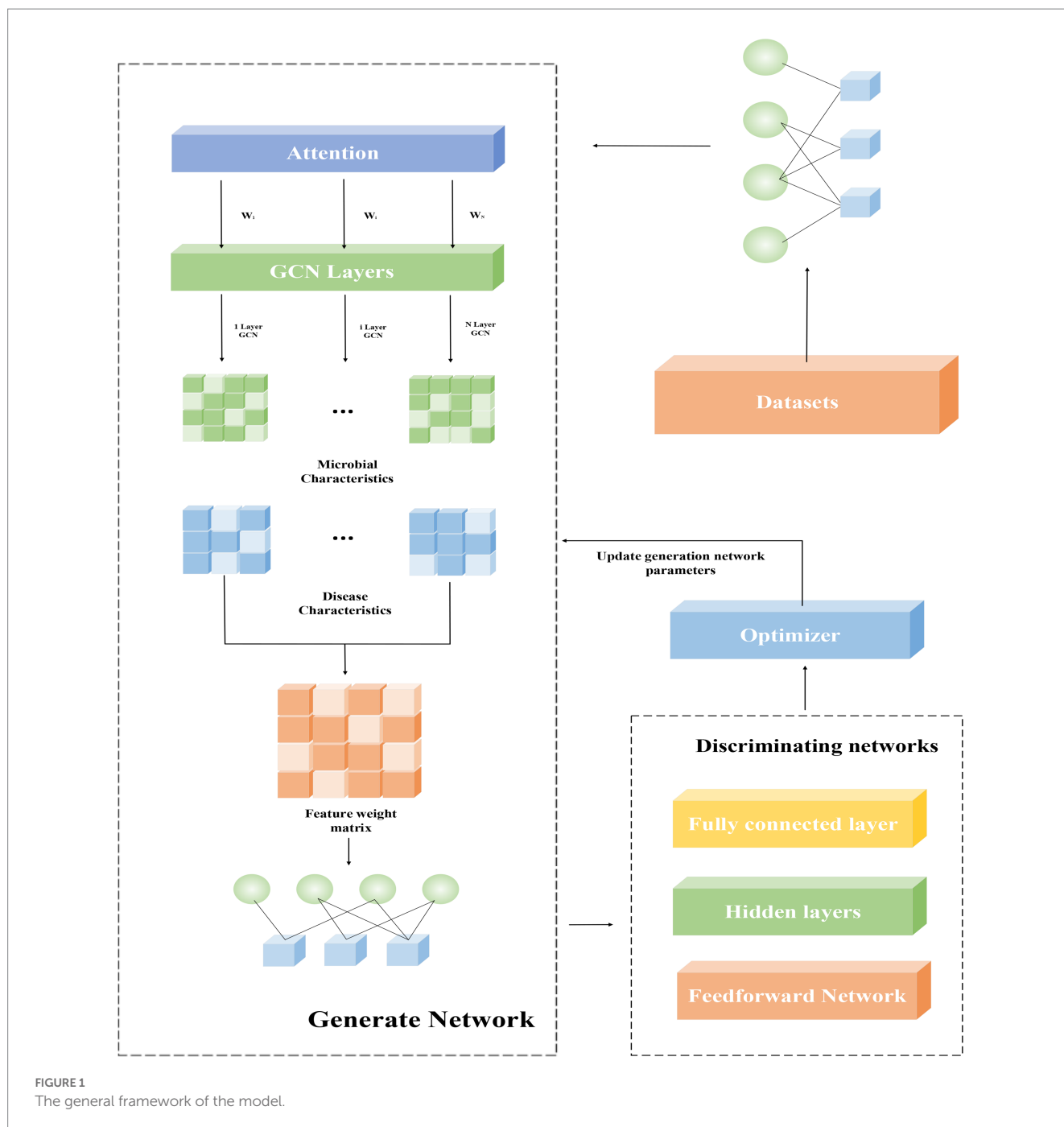
FIGURE 1
The general framework of the model.

The game process of generative adversarial networks can be expressed as follows:

$$\min\max V(D,G) =$$
$$E_{x \sim p_{data}(x)}\left[logD(x)\right] + E_{z \sim p(z)}\left[1 - logD(G(z))\right] \quad (17)$$

Among them, x is the real feature matrix, and $G(z)$ is the feature matrix generated by the generation network. $p_{data}(x)$ is the probability distribution of x, and $p(z)$ is the probability distribution of z. The optimization goal of training D to adjust its parameters is to maximize $D(x)$ and minimize $D(G(x))$, and the optimization goal of training G to adjust its parameters is to

minimize $\max V(D,G)$. E stands for entropy, $x \sim p_{data}(x)$ stands for x is from $p_{data}(x)$ real data distribution. The meaning represented by $E_{x \sim p_{data}(x)}\left[logD(x)\right]$ is the entropy value from the real data distribution after passing the identifier. For data from the real data distribution, the ideal goal of the discriminator is to fully identify it, that is, predict the result as 1. Therefore, $E_{x \sim p_{data}(x)}\left[logD(x)\right]$ can also be regarded as the probability of the discriminator to distinguish real data, and the higher the probability, the better. The log function does not affect the relationship between variables, and its function is to amplify our loss to facilitate the calculation and optimization of the model. $E_{z \sim p(z)}\left[1 - logD(G(z))\right]$ can be regarded as the entropy value after the input generated data passes through the discriminator, and also represents the probability

of the discriminator to distinguish the fake sample data. The smaller the probability, the better. $\min \max V(D,G)$ is expressed as a confrontation between the generator and the discriminator. The generator $G(\bullet)$ hopes that the discriminator cannot distinguish fake samples, so it hopes to minimize the result of $1 - logD(G(z))$. The discriminator is the opposite, it hopes to better distinguish between true and false, that is, the result of maximizing $1 - logD(G(z))$. This is also the origin of this formula. At the end of training, there will often be a balanced form.

The core of the principle of generative adversarial networks lies in the game between the generative network and the decision network. The core of the generative network is composed of GCN layers. In order to deepen the model depth of the generative network and thus generate more accurate prediction results, we use a residual network-like idea to optimize the model. We deepen the network while retaining the shallow features according to the weights, which makes the model less susceptible to phenomena such as oversmoothing and gradient explosion during the iterative process. As shown in Figure 2, the direct mapping is shown on the left, and the associated graph convolution operation and activation function are shown on the right.

The purpose of adding this structure is to increase the depth of the network. Under this premise, problems such as over-smoothing and gradient explosion are avoided. At the same time, combined with the attention mechanism, we have carried out weight ratios on both sides on the basis of similar residual ideas to achieve better results. Its formula derivation is as follows:

$$h_l = h_0 + \sum_{i,j=1}^{L} F(h_i, W_j) \quad (18)$$

Among them, h_L is the feature matrix output by each layer, and $l \in \{1,..,L\}$. W_j is the weight assigned to each layer, and $F(\bullet)$ is the graph convolution function.

And the relevant formula of $F(\bullet)$ is as follows:

$$F(z,W)_l = f\left(F(z)_{l-1}, Y\right) = \mu\left(D^{-\frac{1}{2}}YD^{-\frac{1}{2}}F(z)_{l-1}W_{l-1}\right) \quad (19)$$

Where $l \in \{1,..,L\}$, $F(z)_l$ is the feature matrix generated by the lth layer GCN network, $D = diag\left(\sum_{j=1}^{N_m+N_d} Y_{i,j}\right)$ is a diagonal matrix, and W_l is the weight matrix trained on the lth layer. And $\mu(\bullet)$ is an activation function. In this paper, the RELU function is used as the activation function. The formula is as follows:

$$RELU(x) = \begin{cases} x, x > 0 \\ 0, x \leq 0 \end{cases} \quad (20)$$

The weight calculation formula of W_l is as follows:

$$W_l = \frac{1}{L} \quad (21)$$

Graph Convolution (GCN) is a convolutional model applied by CNN in the field of graph structure. Different from CNN to achieve feature extraction by processing pixels, graph convolution uses spectral graph theory to map the graph structure transformation to the frequency domain through Fourier transform for processing, and finally perform inverse transformation. Compared with CNN that handles neat pixels, GCN can more effectively extract the correlation features between two points. For data with associated structures, the ability to effectively extract spatial features brought by GCN can better help them complete their tasks. In our model, the reconstructed heterogeneous network feature matrix is input into the generative network and processed as the input of the GCN model. Formula (19) reflects the training process of the GCN model, and z is the input data. The function of $D^{-\frac{1}{2}}YD^{-\frac{1}{2}}$ is to dilute the importance of nodes with high degrees, and to balance the weight information of nodes with different degrees. Therefore, formula (19) can also be simplified as:

$$F(z,W)_l = \mu\left(\tilde{Y}F(z)_{l-1}W_{l-1}\right) \quad (22)$$

Among them, the role of $\tilde{Y}F(z)_{l-1}$ is to retain the information inherited by the upper layer nodes during the information transmission process, that is, to aggregate the information of the surrounding nodes to update the information of its own nodes.

The role of the discriminator is to distinguish between real and fake samples, and our discriminator consists of a fully connected feed-forward network, a hidden layer and an output layer. The discriminator alternately receives generated samples and real samples, and updates the parameters of the generated network through the discriminative results. Here we adopt the framework of WassersteinGAN to train the discriminator. The biggest difference between WGAN and traditional GAN is that the output layer is a linear layer and does not require a nonlinear activation function. Expressed in a formula it is:

$$D(z) = \mu(z'W_h + b_h)W_o + b_o \quad (23)$$

Among them, z is the input data, and z is the long vector after dimension reconstruction. $\mu(\bullet)$ is the activation function of the hidden layer, W_h and b_h are the hidden layer parameters, and W_o and b_o are the output layer parameters.

As shown in Algorithm 1, the input is a known microbial-disease association matrix A. The similarity matrix of microorganisms and diseases is computed to construct the heterogeneous network Y. The new feature matrix is fed into the generative network. After initializing the optimizer, the generated prediction results are output after N rounds of training. The generated prediction results and sample data are input into the discriminator, and the parameter information of the generative network is updated according to the output results of the discriminator, and the completed generative network model is saved after several rounds of training.

Algorithm 1: Algorithm of our proposed method

Inputs: Known associations matrix $A \in \mathbb{R}^{n_m \times n_d}$, microbe similarity matrix $K_S^m \in \mathbb{R}^{N_m \times N_m}$, disease similarity matrix $K_S^d \in \mathbb{R}^{n_d \times n_d}$;

Output: The completed training of the generative network model

Step 1: Constructing the heterogeneous network $Y \in \mathbb{R}^{(n_m + n_d) \times (n_d + n_m)}$ according to Formula (16);

Step 2: Input the feature matrix into the generative network, initializing Optimizer Parameter Information;

Step 3: for $i = 1 \rightarrow N$ do (N is the number of training rounds of the generative adversarial network)

    for $l = 1 \rightarrow L$ do (L is the depth of the graph convolution model)

    Compute the feature embedding of the L layer and output the generated prediction results

    end for

    Input the generated results and sample data into the decision network

    Update optimizer parameter information

    end for

Step 4: Save the model of the generative network

4. Experiments and results

4.1. Experimental setup

In this section, we adopted 5-fold cross validation(5cv) and 2-fold cross validation to assess the performance of our model. In the k-fold cross validation framework, all known microbe-disease associations in HMDAD and Disbiome were divided to k-subsets. In the process of model training, $(k-1)$-subsets are selected as the training set, and the remaining one as the test set. It is worth noting that there are no known negative samples, we regarded unknown associations as negative samples. After the training samples are input into MADGAN, all association pairs will get a predictive value. If the prediction score is higher than the given threshold, it will be considered as successful prediction. Obviously, different true positive rate and false positive rate can be obtained when setting different thresholds. The specific calculation formula is as follows:

$$TPR = \frac{TP}{TP + FN}$$

$$FPR = \frac{FP}{FP + TN} \qquad (24)$$

Where TP and TN represent the numbers of positive samples correctly judged as positive samples and negative samples correctly judged as negative samples, respectively; FP and FN are the numbers of negative samples incorrectly judged as positive samples and positive samples incorrectly judged as negative samples. By setting different thresholds, we can get multiple groups of different TPRs and FPRs. Then, TPR and FPR under different thresholds are taken as the x-axis and y-axis respectively, the receiver operating characteristics (ROC) can be further plotted, and the area under the line is taken to evaluate the prediction performance of the model.

4.2. Parameter analysis

We performed multiple experimental and parametric analyses on the HMDAD database and the Disbiome database, respectively. As shown in Figure 3, we analyzed the experimental results generated by HMDAD in terms of the number of layers and embedding. We used a similar idea of residual network to deepen the number of layers of GCN to 4. After several rounds of training, the experimental results and loss values were maintained at a certain level, but we could see from the experimental results that after the number of layers was raised to 5, the experimental results could not be maintained at a certain level as in the previous layers, which we judged to be due to the limitation of the size of the dataset that made it impossible to deepen the network further. We judge that this is due to the limitation of the dataset size, which makes it impossible to deepen the network further, otherwise the phenomenon of oversmoothing will occur. We also compared different embedding values. Different embedding values take different time to train. When the embedding value is 128, the training time cost is greater than when the embedding value is 32. However, when the model depth is deepened to 5 layers, the embedding value of 128 cannot maintain good experimental results, and the embedding values of 32 and 64 are not affected much, but we think that further deepening the model depth and embedding values of 32 and 64 is also oversmoothing can occur, resulting in poor results.

For the Disbiome database, we also conducted multiple experiments, but the Disbiome database is much larger than the HMDAD database, and we were able to maintain the results at a certain level after deepening the GCN layers with our network up to 20 layers, without reaching the limit. We did not find the limit value due to the limitation of the experimental equipment, but we can understand that the experimental results did not deteriorate after deepening to more than 20 layers.

4.3. Comparison with state-of-the-art methods

In order to evaluate the performance of MADGAN, we compare our model with six state-of-the-art methods that includes network-based methods, binary local features-based methods, matrix factorization/completion-based methods and graph neural network-based methods. KATZHMDA and NTSHMDA are network-based methods, NGRHMDA and BiRWMP are binary local features-based methods, GRNMFHMDA is matrix factorization-based method, and GATMDA is graph neural network-based method. The comparison results of all these methods were shown in Tables 1, 2 respectively.

As shown in Tables 1, 2, we used 5 times of cross-validation and 2 times of cross-validation to conduct comparative experiments on the two databases. In experiments on the HMDAD database, our model performs better than other models. The 5-fold cross-validation method makes better use of the data set than the 2-fold cross-validation method, so it performs better. The data sample size of the Disbiome database is much larger than that of HMDAD, and its training time is also much longer than that of HMDAD. However, compared with HMDAD, the experimental results of all models have declined. We believe that part of the reason is that the depth of the model cannot support the training of a large number of sample data. Even if we use
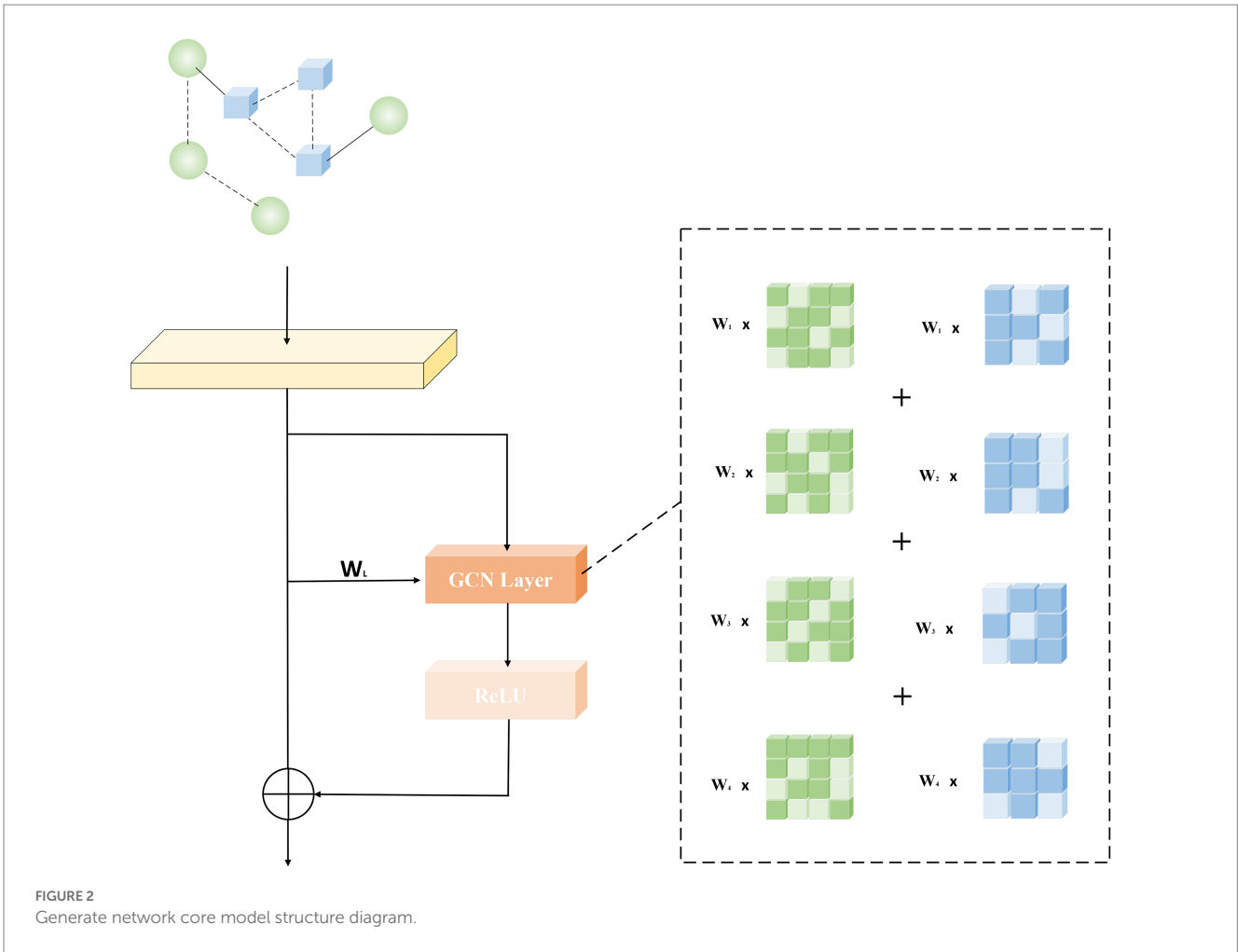
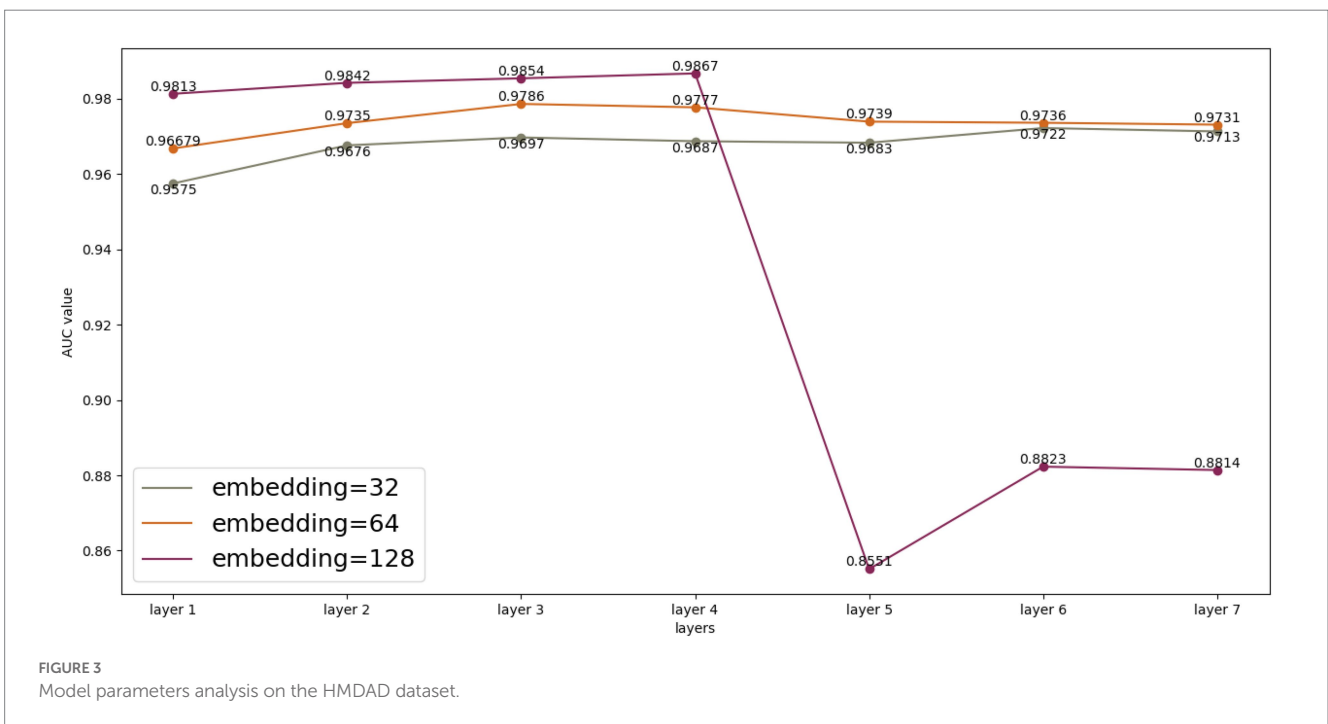**FIGURE 2**
Generate network core model structure diagram.

FIGURE 3
Model parameters analysis on the HMDAD dataset.

TABLE 1 Comparison performance between our model and state-of-the-art models based on HMDAD dataset.

Methods	AUC(5-fold cv)	AUC(2-fold cv)
KATZHMDA (Zhu et al., 2021) (network-based)	0.8703±0.0199	0.8755±0.0103
NTSHMDA (Luo and Long, 2018) (network-based)	0.8982±0.0312	0.8615±0.0151
NGRHMDA (Huang et al., 2017) (binary local features-based)	0.8921±0.0327	0.8929±0.0059
BiRWMP (Luo and Xiao, 2017) (binary local features-based)	0.8777±0.0089	0.8698±0.0079
GRNMFHMDA (He et al., 2018) (matrix factorization-based)	0.8806±0.0156	0.8756±0.0164
GATMDA (Long et al., 2021) (graph neural network-based)	0.9554±0.0184	0.9538±0.0049
Our model	0.9867±0.0078	0.9708±0.0117

TABLE 2 Comparison performance between our model and state-of-the-art models based on Disbiome dataset.

Methods	AUC(5-fold cv)	AUC(2-fold cv)
KATZHMDA (Zhu et al., 2021) (network-based)	0.6779±0.0141	0.6696±0.0058
NTSHMDA (Luo and Long, 2018) (network-based)	0.8294±0.0071	0.8086±0.0058
NGRHMDA (Huang et al., 2017) (binary local features-based)	0.8313±0.0052	0.8233±0.0046
BiRWMP (Luo and Xiao, 2017) (binary local features-based)	0.8344±0.0089	0.8139±0.0060
GRNMFHMDA (He et al., 2018) (matrix factorization-based)	0.8609±0.0047	0.8501±0.0017
GATMDA (Long et al., 2021) (graph neural network-based)	0.9307±0.0079	0.9296±0.0154
Our model	0.9428±0.0026	0.9290±0.0068

the method to deepen the depth of the model, it can only slightly improve the experimental effect. Another part of the reason may be because of the equipment environment.

5. Case study

In this section, we choose three diseases of asthma, Chronic Obstructive Pulmonary Disease (COPD) and Type 2 Diabetes (T2D) for case studies on the HMDAD to further verify the performance of

our model. Specifically, we rank the above three related microorganisms in the predicted score results, and then select the top 20 microorganisms and evaluate the prediction performance of MADGAN through literature retrieval.

Asthma is a disease with heterogeneous process, accompanied by recurrent wheezing, chest tightness, dyspnea, indirect cough and other symptoms(Al-Moamary et al., 2021). It is reported that in 2010, about 8% of people were affected by asthma, especially in children, and the incidence rate is still rising(Guilbert et al., 2014). Asthma has been proved to be closely related to microorganisms(Çalışkan et al., 2013). For example, Haemophilia, Neisseria and Moraxella in the lungs of asthmatic patients have been proved to be closely related to the increased risk of neonatal oral and pharyngeal asthma, and Staphylococcus has been found in the respiratory tract of asthmatic children(Sullivan et al., 2016). These findings may provide a new method for the treatment of asthma. We choose the top 20 microorganisms related to asthma predicted by our model and then search the literature for further verification. The results are shown in the Table 3.

COPD is a lung disease that worsens over time, as long as the symptoms are shortness of breath and cough. By 2015, COPD patients accounted for about 2.4% of the global population (James et al., 2018). Due to the high smoking rate and aging population in developing countries, the death toll of COPD patients is rising rapidly. Although the treatment can delay the deterioration of COPD, there is no cure. Considering that there is a lot of evidence indicating the association between microbiome and COPD, for example, Galiana et al. (2014) found that the diversity of patients with high COPD was lower than that of patients with mild and moderate COPD. Therefore, we select the top 20 microorganisms related to COPD predicted by our model and then search the literature for further verification. The results are shown in the Table 4.

6. Conclusion

Deeply understanding the relationship between microorganisms and diseases can not only reveal the pathogenesis of more human diseases, but also provide new insights into disease prevention, diagnosis and treatment, thus promoting human health. Predicting the potential microbe-disease associations can help biologists to screen the most relevant microorganisms that cause diseases, thus reducing the time and cost of biological verification experiment (Zhou et al., 2017; Uchiyama et al., 2019). In this paper, we developed a deep learning model, named MADGAN, to predict potential microbe-disease associations. We adequately exploit multi-sources of abundant biological data to capture similarity features of microbes and diseases. This helps to predict new microbes (or new diseases) with few or no known association. In order to derive more informative representations, we propose graph convoluted neural network to learn representations for microbes and diseases. Meanwhile, the model is trained through the game between the generation network and the decision network. Finally, we utilized residual network and the cross-level weight distribution structure to enhance the depth of the network to prevent over-smoothing during model training. Comprehensive experiments demonstrated that MADGAN achieved satisfactory predictive performance.

However, although our model has good prediction performance, it still has some limitations and is expected to be further improved in the future. On the one hand, our model is a supervised learning framework, which means that our model cannot predict all new microorganisms

TABLE 3 The top 20 asthma-associated microbes predicted by MADGAN.

Rank	Microbe	Evidence
1	*Clostridium innocuum*	PMID:18672296
2	*Staphylococcus epidermidis*	PMID:6694502
3	*Streptobacillus*	PMID:6326694
4	*Burkholderiales bacterium* Smarlab 3,302,047	Unconfirmed
5	*Dorea*	PMID:30937143
6	*Stenotrophomonas maltophilia*	PMID:20537287
7	*Mannheimia*	PMID:10967288
8	Rikenellaceae	PMID:33204702
9	*Streptococcus parasanguinis*	PMID:17950502
10	*Yersinia*	PMID:10719781
11	*Alistipes*	PMID:33759390
12	*Corynebacterium*	PMID:22994424
13	*Erysipelotrichales*	PMID:22994424
14	*Mobiluncus*	Unconfirmed
15	*Cronobacter*	Unconfirmed
17	Eubacteriaceae	Unconfirmed
18	Unidentified bacterium ZF3	Unconfirmed
19	*Prevotellaceae*	PMID: 34422359
20	*Oxalobacteraceae*	PMID: 21194740

TABLE 4 The top 20 COPD-associated microbes predicted by MADGAN.

Rank	Microbe	Evidence
1	*Bacteroides*	PMID: 36498063
2	*Bacteroides* sp. CJ78	Unconfirmed
3	*Bacteroides vulgatus*	Unconfirmed
4	*Bacteroidetes*	PMID: 33063421
5	*Clostridiales bacterium* 80/3	Unconfirmed
6	*Clostridium cocleatum*	PMID:20857523
7	*Clostridium ramosum*	Unconfirmed
8	*Enterococcus*	PMID:24629344
9	*Erwinia*	Unconfirmed
10	*Escherichia*	PMID: 21605476
11	*Eubacteriaceae*	Unconfirmed
12	*Firmicutes*	PMID: 32353489
13	*Firmicutes bacterium* EG14	Unconfirmed
14	*Fusobacterium*	PMID: 35034433
15	*Verrucomicrobia*	PMID: 32295442
17	*Actinomyces*	PMID: 31174538
18	*Lachnospiraceae bacterium* A2	Unconfirmed
19	*Enterococcus faecalis*	PMID: 26623628
20	*Clostridia bacterium* TSW07CA7	Unconfirmeda

and diseases. In the future, we will consider integrating multiple prior biological information, such as microbe-drug disease association and drug-disease association, to develop an unsupervised learning framework. On the other hand, it is still a huge challenge for MADGAN to forecast on large-scale datasets. In the future, we will consider integrating the results of multiple datasets to build datasets, so as to improve the prediction performance of the model on large datasets.

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 authors.

Author contributions

WH and XY produced the main ideas, and did the modeling, computation and analysis and also wrote the manuscript. LW and XZ provided supervision and effective scientific advice and related ideas, research design guidance, and added value to the article through editing and contributing completions. All authors contributed to the article and approved the submitted version.

Funding

Acknowledgments

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

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2023.1159076/full#supplementary-material

References

Al-Moamary, M. S., Alhaider, S. A., Alangari, A. A., Idrees, M. M., Zeitouni, M. O., Al Ghobain, M. O., et al. (2021). The Saudi initiative for asthma-2021 update: guidelines for the diagnosis and management of asthma in adults and children. *Ann. Thorac. Med.* 16, 4–56. doi: 10.4103/atm.ATM_697_20

Arjovsky, M., Chintala, S., and Bottou, L.. (2017). *Wasserstein Generative Adversarial Networks*. International Conference on Machine Learning, pp. 214–223.

Çalışkan, M., Bochkov, Y. A., Kreiner-Møller, E., Bønnelykke, K., Stein, M. M., Du, G., et al. (2013). Rhinovirus wheezing illness and genetic risk of childhood-onset asthma. *N. Engl. J. Med.* 368, 1398–1407. doi: 10.1056/NEJMoa1211592

Cenit, M. C., Sanz, Y., and Codoñer-Franch, P. (2017). Influence of gut microbiota on neuropsychiatric disorders. *WJG* 23, 5486–5498. doi: 10.3748/wjg.v23.i30.5486

Chen, X., Huang, Y.-A., You, Z.-H., Yan, G. Y., and Wang, X. S. (2017). A novel approach based on KATZ measure to predict associations of human microbiota with non-infectious diseases. *Bioinformatics* 33, 733–739. doi: 10.1093/bioinformatics/btw715

Cheng, Y., Gong, Y., Liu, Y., Song, B., and Zou, Q. (2021). Molecular design in drug discovery: a comprehensive review of deep generative models. *Brief. Bioinform.* 22:bbab344. doi: 10.1093/bib/bbab344

Cryan, J. F., and Dinan, T. G. (2012). Mind-altering microorganisms: the impact of the gut microbiota on brain and behaviour. *Nat. Rev. Neurosci.* 13, 701–712. doi: 10.1038/nrn3346

Dai, H., Chen, C., Li, Y., and Yuan, Y. (2021). GCNGAN: translating natural language to programming language based on GAN. *J. Phys.* 1873:012070. doi: 10.1088/1742-6596/1873/1/012070

Desbonnet, L., Garrett, L., Clarke, G., Kiely, B., Cryan, J. F., and Dinan, T. G. (2010). Effects of the probiotic *Bifidobacterium infantis* in the maternal separation model of depression. *Neuroscience* 170, 1179–1188. doi: 10.1016/j.neuroscience.2010.08.005

Galiana, A., Aguirre, E., Rodriguez, J. C., Mira, A., Santibanez, M., Candela, I., et al. (2014). Sputum microbiota in moderate versus severe patients with COPD. *Eur. Respir. J.* 43, 1787–1790. doi: 10.1183/09031936.00191513

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., et al. (2020). Generative adversarial networks. *Commun. ACM* 63, 139–144. doi: 10.1145/3422622

Guarner, F., and Malagelada, J.-R. (2003). Gut flora in health and disease. *Lancet* 361, 512–519. doi: 10.1016/S0140-6736(03)12489-0

Guilbert, T. W., Mauger, D. T., and Lemanske, R. F. (2014). Childhood asthma-predictive phenotype. The journal of allergy and clinical immunology. *In Pract.* 2, 664–670. doi: 10.1016/j.jaip.2014.09.010

He, B. S., Peng, L. H., 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, Y. J. (2013). Asthma microbiome studies and the potential for new therapeutic strategies. *Curr Allergy Asthma Rep* 13, 453–461. doi: 10.1007/s11882-013-0355-y

Huang, Y.-A., You, Z.-H., Chen, X., Huang, Z. A., Zhang, S., and Yan, G. Y. (2017). Prediction of microbe–disease association from the integration of neighbor and graph with collaborative recommendation model. *J. Transl. Med.* 15, 1–11. doi: 10.1186/s12967-017-1304-7

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

Integrative HMP (iHMP) Research Network Consortium (2014). The integrative human microbiome project: dynamic analysis of microbiome-host omics profiles during periods of human health and disease. *Cell Host Microbe* 16, 276–289. doi: 10.1016/j.chom.2014.08.014

James, S. L., Abate, D., Abate, K. H., Abay, S. M., Abbafati, C., Abbasi, N., et al. (2018). Global, regional, and national incidence, prevalence, and years lived with disability for 354 diseases and injuries for 195 countries and territories, 1990–2017: a systematic analysis for the global burden of disease study 2017. *Lancet* 392, 1789–1858. doi: 10.1016/S0140-6736(18)32279-7

Janssens, Y., Nielandt, J., Bronselaer, A., Debunne, N., Verbeke, F., Wynendaele, E., et al. (2018). Disbiome database: linking the microbiome to disease. *BMC Microbiol.* 18:50. doi: 10.1186/s12866-018-1197-5

Karras, T., Laine, S., and Aila, T.. (2019). *A Style-Based Generator Architecture for Generative Adversarial Networks*. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4401–4410.

Kau, A. L., Ahern, P. P., Griffin, N. W., Goodman, A. L., and Gordon, J. I. (2011). Human nutrition, the gut microbiome and the immune system. *Nature* 474, 327–336. doi: 10.1038/nature10213

Kim, N., Yun, M., Oh, Y. J., and Choi, H. J. (2018). Mind-altering with the gut: modulation of the gut-brain axis with probiotics. *J. Microbiol.* 56, 172–182. doi: 10.1007/s12275-018-8032-4

Lei, K., Qin, M., Bai, B., Zhang, G., and Yang, M.. (2019). *GCN-GAN: A Non-linear Temporal Link Prediction Model for Weighted Dynamic Networks*. IEEE INFOCOM 2019-IEEE Conference on Computer Communications. IEEE, pp. 388–396.

Li, X., Watanabe, K., and Kimura, I. (2017). Gut microbiota Dysbiosis drives and implies novel therapeutic strategies for diabetes mellitus and related metabolic diseases. *Front. Immunol.* 8:1882. doi: 10.3389/fimmu.2017.01882

Long, Y., Luo, J., Zhang, Y., and Xia, Y. (2021). Predicting human microbe–disease associations via graph attention networks with inductive matrix completion. *Brief. Bioinform.* 22:bbaa146. doi: 10.1093/bib/bbaa146

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

Luo, J., and Xiao, Q. (2017). A novel approach for predicting microRNA-disease associations by unbalanced bi-random walk on heterogeneous network. *J. Biomed. Inform.* 66, 194–203. doi: 10.1016/j.jbi.2017.01.008

Ma, W., Zhang, L., Zeng, P., Huang, C., Li, J., Geng, B., et al. (2017). An analysis of human microbe–disease associations. *Brief. Bioinform.* 18, 85–97. doi: 10.1093/bib/bbw005

Quigley, E. M. M. (2013). Gut bacteria in health and disease. *Gastroenterol. Hepatol.* 9, 560–569.

Schwabe, R. F., and Jobin, C. (2013). The microbiome and cancer. *Nat. Rev. Cancer* 13, 800–812. doi: 10.1038/nrc3610

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, Z., Jiang, Z., and Bao, W. (2017). CMFHMDA: collaborative matrix factorization for human microbe-disease association prediction. *Intell. Comput. Theor. Appl.*, 261–269. doi: 10.1007/978-3-319-63312-1_24

Sullivan, A., Hunt, E., MacSharry, J., and Murphy, D. M. (2016). The microbiome and the pathophysiology of asthma. *Respir. Res.* 17:163. doi: 10.1186/s12931-016-0479-4

Uchiyama, I., Mihara, M., Nishide, H., Chiba, H., and Kato, M. (2019). MBGD update 2018: microbial genome database based on hierarchical orthology relations covering closely related and distantly related comparisons. *Nucleic Acids Res.* 47, D382–D389. doi: 10.1093/nar/gky1054

Wang, D., Wang, J., Lu, M., Song, F., and Cui, Q. (2010). Inferring the human microRNA functional similarity and functional network based on microRNA-associated diseases. *Bioinformatics* 26, 1644–1650. doi: 10.1093/bioinformatics/btq241

Wei, H., and Liu, B. (2020). iCircDA-MF: identification of circRNA-disease associations based on matrix factorization. *Brief. Bioinform.* 21, 1356–1367. doi: 10.1093/bib/bbz057

Wu, H., Feng, J., Tian, X., Xu, F., Liu, Y., Wang, X. F., et al. (2019). *secGAN: A Cycle-Consistent GAN for Securely-recoverable Video Transformation*. Proceedings of the 2019 Workshop on Hot Topics in Video Analytics and Intelligent Edges, pp. 33–38.

Xu, J., and Li, Y. (2006). Discovering disease-genes by topological features in human protein–protein interaction network. *Bioinformatics* 22, 2800–2805. doi: 10.1093/bioinformatics/btl467

Zeng, X., Tu, X., Liu, Y., Fu, X., and Su, Y. (2022). Toward better drug discovery with knowledge graph. *Curr. Opin. Struct. Biol.* 72, 114–126. doi: 10.1016/j.sbi.2021.09.003

Zhang, W., Yang, W., Lu, X., Huang, F., and Luo, F. (2018). The bi-direction similarity integration method for predicting microbe-disease associations. *IEEE Access* 6, 38052–38061. doi: 10.1109/ACCESS.2018.2851751

Zheng, H., Li, X., Li, Y., Yan, Z., and Li, T. (2022). GCN-GAN: integrating graph convolutional network and generative adversarial network for traffic flow prediction. *IEEE Access* 10, 94051–94062. doi: 10.1109/ACCESS.2022.3204036

Zhou, T., Tan, L., Cederquist, G. Y., Fan, Y., Hartley, B. J., Mukherjee, S., et al. (2017). High-content screening in hPSC-neural progenitors identifies drug candidates that inhibit Zika virus infection in fetal-like organoids and adult brain. *Cell Stem Cell* 21, 274–283.e5. doi: 10.1016/j.stem.2017.06.017

Zhu, L., Duan, G., Yan, C., and Wang, J. (2021). Prediction of microbe-drug associations based on chemical structures and the KATZ measure. *Curr. Bioinforma.* 16, 807–819. doi: 10.2174/1574893616666210204144721

Zhu, J. Y., Park, T., Isola, P., and Efros, A. A.. (2017). *Unpaired Image-to-image Translation Using Cycle-consistent Adversarial Networks*. Proceedings of the IEEE International Conference on Computer Vision, pp. 2223–2232.