*Briefings in Bioinformatics*, 22(3), 2021, 1–13

**doi: 10.1093/bib/bbaa146**
Problem Solving Protocol

# Predicting human microbe–disease associations via graph attention networks with inductive matrix completion

Yahui Long, Jiawei Luo, Yu Zhang and Yan Xia

Corresponding author: Jiawei Luo, College of Computer Science and Electronic Engineering, Hunan University, Changsha 410000, China. Tel.:+86-0731-88821971; E-mail: luojiawei@hnu.edu.cn



## Abstract

**Motivation:** human microbes play a critical role in an extensive range of complex human diseases and become a new target in precision medicine. *In silico* methods of identifying microbe–disease associations not only can provide a deep insight into understanding the pathogenic mechanism of complex human diseases but also assist pharmacologists to screen candidate targets for drug development. However, the majority of existing approaches are based on linear models or label propagation, which suffers from limitations in capturing nonlinear associations between microbes and diseases. Besides, it is still a great challenge for most previous methods to make predictions for new diseases (or new microbes) with few or without any observed associations.
**Results:** in this work, we construct features for microbes and diseases by fully exploiting multiply sources of biomedical data, and then propose a novel deep learning framework of graph attention networks with inductive matrix completion for human microbe-disease association prediction, named GATMDA. To our knowledge, this is the first attempt to leverage graph attention networks for this important task. In particular, we develop an optimized graph attention network with talking-heads to learn representations for nodes (i.e. microbes and diseases). To focus on more important neighbours and filter out noises, we further design a bi-interaction aggregator to enforce representation aggregation of similar neighbours. In addition, we combine inductive matrix completion to reconstruct microbe-disease associations to capture the complicated associations between diseases and microbes. Comprehensive experiments on two data sets (i.e. HMDAD and Disbiome) demonstrated that our proposed model consistently outperformed baseline methods. Case studies on two diseases, i.e. asthma and inflammatory bowel disease, further confirmed the effectiveness of our proposed model of GATMDA.
**Availability:** python codes and data set are available at: https://github.com/yahuilong/GATMDA.
**Contact:** luojiawei@hnu.edu.cn.

**Key words:** microbe–disease associations; deep learning; graph attention networks; matrix completion.

**Yahui Long** is a PhD candidate at Hunan University, China and also a joint PhD candidate at Nanyang Technological University, Singapore. His research interests include deep learning and bioinformatics.
**Jiawei Luo** is a full professor at Hunan University, China. Her research interests include graph theory, data mining and bioinformatics.
**Yu Zhang** is a PhD candidate at Nanyang Technological University, Singapore. Her research interests include bioinformatics and deep learning.
**Yan Xia** is a lecturer at Hunan University, China. Her research interests include big data, machine learning and bioinformatics.
**Submitted:** 23 April 2020; **Received (in revised form):** 7 June 2020

## Introduction

Microbe or microorganism is a category of the microscopic living organism, which may exist in its single-celled form or colony cells [1]. Accumulated researches have demonstrated that a large number of microbe communities, mainly consisting of viruses, archaea, bacteria and protozoa, are shown to closely interact with human hosts [2, 3]. In general, they reside in and on various human body organs, such as an oral cavity, gut, skin and lung, with the majority residing in the gastrointestinal tract [4]. In fact, most of the human commensal microbial communities are harmless to human health and even have a mutualistic relationship with their human hosts. Particularly, human microbes are commonly considered as 'forgotten organ' of human beings owing to its similar metabolic capacity to the liver, such as promoting nutrient absorption, providing protection from pathogens and strengthening metabolic capability [5]. Therefore, the dysbiosis or imbalance of microbial communities could cause human diseases, such as liver diseases [6], diabetes [7], asthma [8] and even cancer [9]. However, although many studies have been conducted to increasingly reveal the roles of microbes in the pathogenic mechanism of human diseases, a systematical understanding of how microbes living in human bodies influence human healths and cause diseases remains poorly known.

Identifying microbe–disease associations not only can help reveal the complex disease-causing mechanism but also provide potential biomarkers for disease diagnosis and prognosis. As the conventional wet lab methods are time consuming, labor intensive and expensive, *in silico* methods are a great complementary and can guide these experiments. Very recently, much attention has been devoted to developing computational methods to predict microbe–disease associations. We can divide these methods into three categories, namely network-based methods, random walk-based methods and matrix factorization/completion-based methods. Network-based methods are the most common ones. For example, Chen *et al.* [10] proposed a KATZ-based model of KATZHMDA for identifying microbe–disease associations based on a heterogeneous network. Huang *et al.* [11] presented a novel computational model of PBHMDA, which implemented a devised depth-first search algorithm on the heterogeneous network to predict microbe–disease associations. Long *et al.* [12] developed a meta-graph method named WMGHMDA, which attempted to infer potential microbe–disease pairs by searching the meta-graphs of the pairs on a heterogeneous network. The random walk has aroused extensive interest in the field of microbe–disease prediction. For instance, Zou *et al.* [13] released a computational framework of BiRWHMDA for microbe–disease prediction, which simultaneously implemented random walk on similarity networks for microbes and diseases. To take network topological similarity into consideration, Luo *et al.* [14] developed an improved random walk-based prediction model of NTSHMDA, which encouraged the walker to select a closer neighbour node. Besides, Yan *et al.* [15] introduced a novel optimized bi-random-based computational model (BRWMDA). Besides, some computational approaches for microbe–disease prediction are developed based on matrix factorization/completion techniques. For example, Shen *et al.* [16] developed a computational method of CMFHMDA, which used collaborative matrix factorization to recover the correlation matrices between diseases and microbes. He *et al.* [17] released a novel matrix factorization-based model named GRNMFHMDA by introducing graph regularization term. Duan *et al.* [18] presented a matrix completion-based model of MCHMDA to uncover microbe–disease associations combining with the fast singular value thresholding algorithm.

Despite the effectiveness of the aforementioned methods for microbe–disease prediction, there are still some limitations to this prediction task. First, network-based methods and random walk-based methods are easily biased towards well-investigated diseases (or microbes). For example, for given diseases (or microbes) that have few known associations, they can obtain insufficiently accurate candidate microbes (or diseases) since sparse links limit information propagation. Second, the majority of previous methods strongly depend on known microbe–disease associations for similarity calculation, which makes these methods unable to achieve prediction when it involves new diseases (or new microbes) due to the absence of training data. Here we define diseases (or microbes) that have no known associations as new diseases (or new microbes). Nowadays, more and more prior biomedical data associated with microbes and diseases are accumulated, such as HMDAD [19], STRING [20] and HumanNet [21], which provides a golden opportunity for us to leverage graph-based deep learning techniques to predict their associations by propagating information from local neighbours to them. Third, matrix factorization/completion can only capture linear associations, which is unable to accurately reflect the nonlinear microbe–disease interaction associations.

To deal with above limitations, deep learning technique is an alternative choice, which has achieved successful applications in various graph-based tasks, such as text classification [22], recommender system [23] and link prediction [24, 25]. In fact, microbe–disease interactions can be represented by a graph/network, where nodes denotes microbes and diseases and interacting microbe–disease pairs form edges in the graph. As such, predicting novel microbe–disease associations can be mapped as a link prediction task in the graph. In particular, the graph attention network (GAT) is a promising deep learning technique due to its great potential in modeling complex graph data, which has drawn increasing attention from different research fields, such as node/graph classification [26, 27], recommender system [28, 29] and semantic segmentation [30]. Recently, graph attention network has been applied for some bioinformatics tasks. For example, Zheng *et al.* [31] proposed a novel graph attention network-based method named GAPDA to identify piRNA (piwi-interacting RNA)-disease associations. Ravindra and Sehanobish [32] presented a graph attention model for inferring disease state from single-cell data. However, so far no work has been done for microbe–disease prediction using GAT. As such, we are motivated to customize GAT for novel microbe–disease prediction. On the other hand, in the link prediction tasks, most deep learning-based models first learn representation vectors for nodes and then treat the inner products of representation vectors as the association probabilities. However, it is still not enough to identify the complex associations between nodes using a simple inner product. Despite the above-mentioned limitation, matrix completion is shown to possess the powerful capability in modeling this complicated interaction associations.

In this work, we propose a novel graph attention network-based framework named GATMDA for microbe–disease association prediction in a bipartite network, combining inductive matrix completion (IMC). First, we derive comprehensive features for microbes and diseases by assembling multiple sources of biological data, such as microbe functional similarity, disease functional similarity and Gaussian kernel similarities. Second, we introduce graph attention networks with talking-heads, a variant of benchmark GAT, to learn node representations, which

enables the model to preserve more informative representations. For each head, we further design a bi-interaction aggregator in the neural embedding/representation aggregation layer to enforce representation aggregation of similar neighbors, such that similar nodes have similar representations. Third, we introduce IMC into the model to reconstruct microbe–disease associations. Experimental results on two data sets, i.e. HMDAD and Disbiome, indicated that our proposed model of GATMDA consistently outperformed seven state-of-the-art methods. Case studies on two common diseases, i.e. asthma and inflammatory bowel disease (IBD), further validated the effectiveness of GATMDA.

Overall, our main contributions are summarized as follows.

- We derived comprehensive features for microbes and diseases by taking full advantage of various biomedical data, such as gene–gene interactions, disease–gene associations and gene-related biological processes. Besides, we sorted out the second microbe–disease association data set Disbiome, which can facilitate future research in this important field.
- We proposed a novel GAT-based framework for predicting microbe–disease associations in a bipartite network. To the best of our knowledge, this is the first work to apply a graph attention network for predicting microbe–disease associations.
- We introduced graph attention network with talking-heads, which encourages information propagation from head to head and thus ensure more informative representations. We further designed a bi-interaction aggregator in the neural representation aggregation layer to strengthen representation aggregations between similar nodes (i.e. microbes and diseases).
- Instead of inner product, we reconstructed microbe–disease associations by adopting IMC, enabling the model to more accurately capture complicated associations between microbes and diseases.
- Our comprehensive experimental results and case studies demonstrated the proposed method of GATMDA outperformed seven state-of-the-art methods significantly on the benchmark HMDAD and Disbiome data sets.

## Materials

### Human microbe–disease associations

We downloaded known microbe–disease associations from database HMDAD (http://www.cuilab.cn/hmdad), which includes 483 experimentally confirmed microbe–disease associations between 39 diseases and 292 microbes [19]. In HMDAD, a microbe–disease pair may include multiply entries from different evidence. Here we consider the same microbe–disease associations from different evidence as a pair. Subsequently, we obtained 450 associations involving 39 diseases and 292 microbes. In addition, Janssens *et al.* [33] has recently released a new microbe–disease association database named Disbiome (https://disbiome.ugent.be/home) where 5573 experimentally confirmed human microbe–disease relationships were collected from previously published literature and different databases about 240 diseases and 1098 microbes. In Disbiome, a microbe–disease pair may be recorded more than one time according to different detection methods. Here we neglect the information of detection methods. After filtering out repetitive data, we finally downloaded 4351 associations between 218 diseases and 1052 microbes. Overall, the statistics of the two microbe–disease association data sets above are shown in Table 1. More details of these two data sets could be found in Supplementary Table R1–R6.

**Table 1.** The statistics for each microbe–disease association data set

| Data sets | # Microbes | # Diseases | # Associations |
|---|---|---|---|
| HMDAD | 292 | 39 | 450 |
| Disbiome | 1052 | 218 | 4351 |

For convenience, we formulated microbe–disease associations as a binary matrix $Y \in \mathbb{R}^{nd \times nm}$ with $nd$ and $nm$ representing the numbers of diseases and microbes, respectively. If there exists an experimentally verified relationship between a disease $d_i$ and a microbe $m_j$, $Y_{ij}$ equals to 1, otherwise 0. Also, we can construct a bipartite network using microbe–disease associations. We define its adjacent matrix $A \in \mathbb{R}^{(nd+nm)\times(nd+nm)}$ as follows:

$$A = \begin{bmatrix} 0 & Y \\ Y^T & 0 \end{bmatrix}. \tag{1}$$

### Microbe functional similarity

In this work, we calculate microbe functional similarity using a similar method to that in [34]. We retrieve protein–protein functional interaction network from STRING v11 database (*https://string-db.org*.). More details about the calculation of microbe functional similarity could be found in [34]. We utilize $FS \in \mathbb{R}^{nm \times nm}$ to represent the microbe functional similarity, where $FS(m_i, m_j)$ denotes the similarity between microbe $m_i$ and $m_j$.

### Disease functional similarity

Motivated by the assumption that similar diseases tend to interact with similar genes [35, 36], we calculate disease functional similarity based on the functional associations between disease-related genes. The latest released HumanNet v2.0 database (https://www.inetbio.org/humannet/download.php) is available for effectively accessing gene interactions [21], where each interaction has an associated log-likelihood score (LLS) that evaluates the probability of a functional linkage between genes. For a disease pair $d_i$ and $d_j$, we first derive their related gene sets $G_i = \{g_{i1}, g_{i2}, \cdots, g_{im}\}$ and $G_j = \{g_{j1}, g_{j2}, \cdots, g_{jn}\}$, respectively. $m$ is the number of genes in $G_i$, while $n$ is the number of genes in $G_j$. We define the functional association between a gene $g$ and a gene set $G = \{g_1, g_2, \cdots, g_k\}$ as follows:

$$F_G(g) = \max_{g_i \in G}(FSS(g, g_i)), \tag{2}$$

where $FSS$ represents the functional similarity score between genes, which is defined as follows:

$$FSS(g_i, g_j) = \begin{cases} 1 & \text{if } i = j, \\ LLS'(g_i, g_j) & \text{if } i \neq j, \end{cases} \tag{3}$$

where $LLS'$ is the normalized LLS of genes, which is defined as follows:

$$LLS'(g_i, g_j) = \frac{LLS(g_i, g_j) - LLS_{min}}{LLS_{max} - LLS_{min}}, \tag{4}$$

where $LLS_{max}$ and $LLS_{min}$ denote the maximum LLS and minimum LLS in HumanNet, respectively.

Finally, we formulate the disease functional similarity as:

$$DF(d_i, d_j) = \frac{\sum_{g_t \in G(d_i)} F_{G(d_j)}(g_t) + \sum_{g_t \in G(d_j)} F_{G(d_i)}(g_t)}{m+n}. \quad (5)$$

### Gaussian interaction profile kernel similarity for diseases and microbes

Inspired by the assumption that functionally similar microbes generally show interaction or non-interaction patterns with similar diseases and vice versa [10], we calculate Gaussian kernel similarity for diseases and microbes by utilizing Gaussian kernel function based on known microbe–disease evidence. Since the $i^{th}$ row and the $j^{th}$ column of adjacent matrix $Y$ represent the interactions between disease $d_i$ or microbe $m_j$ and all microbes or all diseases, we represent $IP(d_i)$ and $IP(m_j)$ as the interaction profiles of disease $d_i$ and microbe $m_j$, respectively. The Gaussian kernel similarities between diseases and microbes are defined as follows:

$$GD(d_i, d_j) = \exp(-\lambda_d \|IP(d_i) - IP(d_j)\|^2), \quad (6)$$

$$GM(m_i, m_j) = \exp(-\lambda_m \|IP(m_i) - IP(m_j)\|^2), \quad (7)$$

where $\lambda_d$ and $\lambda_m$ represent the normalized kernel bandwidths and are defined as follows:

$$\lambda_d = \lambda_d^{'} / \left( \frac{1}{nd} \sum_{i=1}^{nd} \|IP(d_i)\|^2 \right), \quad (8)$$

$$\lambda_m = \lambda_m^{'} / \left( \frac{1}{nm} \sum_{i=1}^{nm} \|IP(m_i)\|^2 \right), \quad (9)$$

where $\lambda_d^{'}$ and $\lambda_m^{'}$ are the original bandwidths, and generally both are set to 1.

### Integrated similarities for diseases and microbes

Since not all diseases have known associated genes, if a given disease is lack of related genes, we cannot generate its functional similarity scores between it and other diseases. As such, to complementary and improve similarity for diseases, we define a new disease similarity by integrating Gaussian kernel disease similarity and disease functional similarity. Specifically, if there is a functional similarity between a disease $d_i$ and a disease $d_j$, the integrated similarity between $d_i$ and $d_j$ is defined as the average of $DF$ and $GD$, otherwise it is equal to the Gaussian interaction profile kernel similarity $GD$. The integrated disease similarity $DS \in \mathbb{R}^{nd \times nd}$ is calculated as follows:

$$DS(d_i, d_j) = \begin{cases} \frac{DF(d_i,d_j)+GD(d_i,d_j)}{2} & \text{if } DF(d_i, d_j) \neq 0, \\ GD(d_i, d_j) & \text{otherwise.} \end{cases} \quad (10)$$

Similarly, the integrated microbe similarity $MS \in \mathbb{R}^{nm \times nm}$ is calculated as follows:

$$MS(m_i, m_j) = \begin{cases} FS(m_i, m_j) & \text{if } FS(m_i, m_j) \neq 0, \\ GM(m_i, m_j) & \text{otherwise.} \end{cases} \quad (11)$$

In consistent with the bipartite network in Equation 1, the feature matrix $X \in \mathbb{R}^{(nd+nm)\times(nd+nm)}$ for microbes and diseases is described as follows:

$$X = \begin{bmatrix} DS & 0 \\ 0 & MS \end{bmatrix}. \quad (12)$$

## Methods

In this work, we propose a novel semi-supervised graph attention network (GAT)-based framework named GATMDA to predict novel microbe–disease associations. As shown in the right part of Figure 1, GATMDA includes three main steps. First, we learn representations for microbes and diseases based on GAT with talking-heads. For each head, we further design a bi-interaction aggregator to encourage information propagation between similar nodes in the neural representation aggregation layer. Second, we introduce inductive matrix completion to learn a decoder in a semi-supervised way for identifying association ratings between diseases and microbes. Third, we reconstruct microbe–disease associations based on the association ratings. Next, we introduce the above three steps in detail.

### Graph attention network with talking-heads

Graph attention network (GAT), proposed by [37], aims to learn representations for nodes on the graph by assigning different weights to different neighbours. In particular, GAT leverages multi-head attention mechanism to stabilize the learning process of self-attention. However, while such an operation can effectively avoid the influence of single self-attention, the learned representations are still insufficiently informative. Besides, different heads are completely independent, which fails to take into account the dependency between heads. To deal with these limitations, one solution is to construct the dependency between different heads by propagating information from head to head. We term the correlative multi-head attention mechanism as talking-head attention mechanism (the details would be introduced in Section 3.3). Here we customize graph attention networks with talking-heads to predict novel microbe–disease associations.

### Attentive representation learning layer

After deriving adjacent matrix $A$ in Section 2.1 and feature matrix $X$ in Section 2.5, we can utilize them to learn representations for diseases and microbes based on GAT with talking-heads. Specifically, for a given node, GAT first learns the importance of its neighbours, then fuses the representations of its neighbours according to their attention scores and subsequently updates its representation based on its current representation and the fused representation of neighbours. In particular, the attention score $e_{ij}^k$ for an association pair between disease $d_i$ and microbe $m_j$ is formulated as follows:

$$e_{ij}^k(d_i, m_j) = f(W_t h_i^k, W_t h_j^k), \quad (13)$$

where $f$ denotes a single-layer feed-forward neural network, parameterized by a weight matrix $W_t$ that transforms input features/representations into high-level features for microbes and diseases. $h^k \in \mathbb{R}^l$ denotes the representation of node in the $k$ ($k \in [1, 2, \ldots K]$)-th attention head with $l$ representing the feature dimension. $H^1$ is defined as the initial feature matrix $X$ of nodes. To make attention scores comparable across different nodes, we further normalize the attention scores using the following

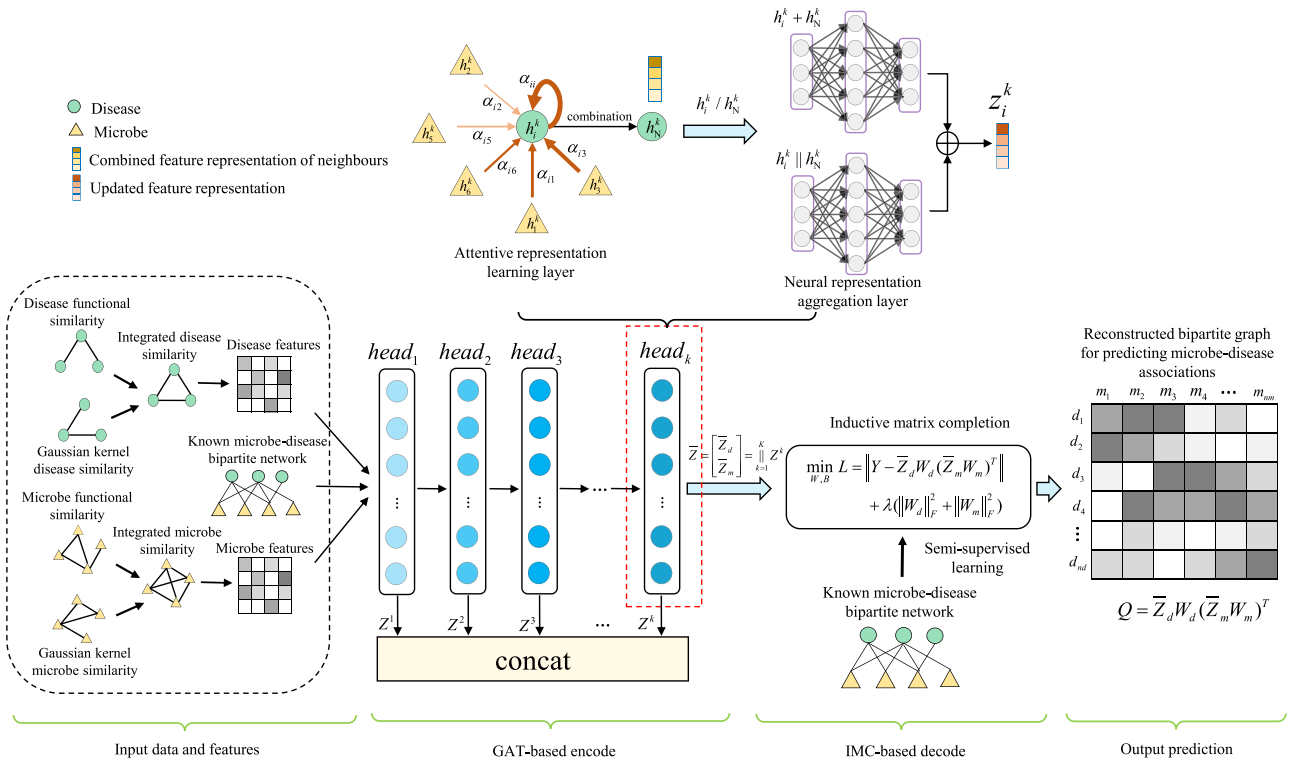


**Figure 1.** The overflow of GATMDA for novel microbe–disease prediction. GATMDA mainly consists of three modules. The first module is to construct input features for microbes and diseases by fully exploiting multiply sources of biomedical data. The second module is designed to learn node representations based on GAT with talking-heads, which includes the attention representation learning layer and the neural representation aggregation layer. The third module aims to reconstruct microbe–disease associations based on IMC in a semi-supervised learning way.

softmax function:

$$\alpha_{ij}^{k} = \frac{\exp(e_{ij}^{k})}{\sum_{t\in\mathcal{N}_i}\exp(e_{it}^{k})}, \tag{14}$$

where $\mathcal{N}_i$ represents the neighbours of disease $d_i$. $\alpha_{ij}^{k}$ means how important microbe $m_j$ would be for disease $d_i$ in the process of information propagation (i.e. the next step).

Considering a disease $d_i$ (or microbe $m_j$), we term the sub-network consisting of itself and neighbour microbes (or diseases) as ego-network. To model the first-order connectivity structure of disease $d_i$, which is determined by adjacent matrix $A$, we calculate the linear combination of its ego-network as follows:

$$h_{\mathcal{N}_i}^{k} = \sum_{t\in\mathcal{N}_i}\alpha_{it}^{k}h_{t}^{k}. \tag{15}$$

## Neural representation aggregation layer

We have obtained the node representation $h^k$ and its ego-network representation $h_{\mathcal{N}}^{k}$. Inspired by the assumption that microbes with similar functions tend to interact with similar diseases [10], we further design a bi-interaction aggregator to aggregate $h^k$ with $h_{\mathcal{N}}^{k}$ based on nonlinear graph neural networks, such that similar nodes are similar in the feature spaces. More specifically, we can update the representation matrix $Z$ of nodes as follows:

$$\begin{aligned} Z^{k} =& LeakyReLU(W_a(h^k + h_{\mathcal{N}}^{k}) + B_a) + \\ & LeakyReLU(W_c(h^k || h_{\mathcal{N}}^{k}) + B_c), \end{aligned} \tag{16}$$

where $W_a \in \mathbb{R}^{l\times r}$, $W_c \in \mathbb{R}^{(l+l)\times r}$, $B_a \in \mathbb{R}^{(nd+nm)\times r}$ and $B_c \in \mathbb{R}^{(nd+nm)\times r}$ are learnable weight and bias matrices, respectively. $r$ represents the transformation size. $\|$ represents concatenation operation and *LeakyReLU* denotes activation function. Here we encode two types of feature interactions between $h^k$ and $h_{\mathcal{N}}^{k}$, which makes the information being propagated sensitive to the relevance between $h^k$ and $h_{\mathcal{N}}^{k}$. In other words, such aggregator encourages to propagate more information between similar nodes.

As mentioned above, to stabilize the learning process of self-attention, benchmark GAT adopts multi-head mechanism. However, in fact, the representations learned in such way are still insufficiently informative due to the independence of different heads. Here we introduce talking-head mechanism to strengthen node representations by constructing dependency between different heads. Specifically, we take the output representation $\overline{Z}^{k-1}$ generated from previous head as the input feature $h^k$ of next head. $h^1$ is defined as the preliminary input feature $x$ of node. This kind of talking-head mechanism enforces information propagated from head to head, which enables the model to incrementally preserve the importance of high-order neighbours. We can thus attain the final representation matrix $\overline{Z} \in \mathbb{R}^{(nd+nm)\times(nd+nm)}$ for diseases and microbes by concatenating the output representations of each head as follows:

$$\overline{Z} = \begin{bmatrix} \overline{Z}_d \\ \overline{Z}_m \end{bmatrix} = \mathop{\Big\|}_{k=1}^{K} Z^{k}. \tag{17}$$

Note that all the above parameters are shared for different heads.

Algorithm 1 GATMDA Algorithm

Input:
adjacency matrix $A \in \mathbb{R}^{(nd+nm)\times(nd+nm)}$; feature matrix $X \in \mathbb{R}^{(nd+nm)\times(nd+k)}$; number of heads $K$; number of neurons $r$; maximum training epoches $T$; learning rate $\eta$.
Output:
probability score matrix Q.

Initialize parameter matrices $W = \left\{W_t, W_a, W_c, W_d, W_m\right\}$ and $B = \left\{B_a, B_c\right\}$ with random values;
$t \longleftarrow 1$;
**repeat**
  **for** $k = 1, 2, \ldots, K$ **do**
    **for** each disease $d_i \in \mathcal{V}_d$ and each microbe $m_j \in \mathcal{V}_m$ **do**
      Computer attention score $\alpha_{it}^k$ with Eq. (14);
      Obtain node ego-network representation $h_{\mathcal{N}_i}^k$:
        $h_{\mathcal{N}_i}^k \longleftarrow \sum_{t\in\mathcal{N}_i} \alpha_{it}^k h_t^k$;
      Update node representation:
        $z^k \longleftarrow LeakyReLU(W_a(h^k + h_{\mathcal{N}_i}^k) + B_a)$
          $+ LeakyReLU(W_c(h^k || h_{\mathcal{N}_i}^k) + B_c)$;
    **end for**
  $h^{k+1} \longleftarrow z^k$; // achieving information propagation between different heads.
  **end for**
  Obtain the final representation for node:
    $\overline{z} \longleftarrow \Big\Vert_{k=1}^{K} z^k$;
  Update $W$ and $B$ by optimizing Eq.(20);
  $t \longleftarrow t + 1$;
**until** $t > T$ or Eq.(20) is converged;
Obtain prediction matrix Q with Eq. (19);
**return** Q.

### Decoder for microbe–disease association reconstruction

We derive the feature matrices $\overline{Z}_d \in \mathbb{R}^{nd\times Kr}$ for diseases and $\overline{Z}_m \in \mathbb{R}^{nm\times Kr}$ for microbes in Equation 17. For previous embedding/representation-based link prediction models, the most common way to infer association pairs is to utilize the inner products of node embeddings/representations to determine their probability scores. However, since the inner product is simple, it is limited to capture the complicated associations between nodes. To handle this challenge, we introduce the inductive matrix completion technique, which is shown to have great potential in modeling association ratings of node pairs, to reconstruct novel disease–microbe associations with the learned representations. The main idea of IMC is to reconstruct a matrix to complete the missing entries based on known entries. We define the loss function in Equation 18 and reconstruct an adjacent matrix $Q \in \mathbb{R}^{nd\times nm}$ for microbe–disease associations in Equation 20:

$$\mathcal{L} = \left\| Y - \overline{Z}_d W_d (\overline{Z}_m W_m)^T \right\|_F^2 + \lambda(\|W_d\|_F^2 + \|W_m\|_F^2), \quad (18)\ (19)$$

$$Q = \overline{Z}_d W_d (\overline{Z}_m W_m)^T, \quad (20)$$

where $W_d$, $W_m$ are trainable weight parameters, and $\lambda$ is decay factor to balance the regularization term.

### Optimization

In this work, GATMDA is trained to learn the parameters by minimizing the loss $\mathcal{L}$ in Equation 18 as follows:

$$\min_{W,B} \mathcal{L}_{\Omega^+\cup\Omega^-} = \left\| Y - \overline{Z}_d W_d (\overline{Z}_m W_m)^T \right\|_F^2 + \lambda(\|W_d\|_F^2 + \|W_m\|_F^2), \quad (21)$$

where $W = \left\{W_t, W_a, W_c, W_d, W_m\right\}$ and $B = \left\{B_a, B_c\right\}$ represent weight and bias matrix sets in our model, respectively. We leverage the Adam optimizer [38] for the optimization. In addition, we adopt negative sampling to better train the model. For each epoch, with the positive samples ($\Omega^+$), we randomly sample equal-size unknown disease–microbe associations as negative samples ($\Omega^-$) for training. Subsequently, we prioritize novel disease–microbe association pairs according to their probability scores calculated by Equation 20. The detailed steps of GATMDA to predict novel microbe–disease associations is described in Algorithm 1.

## Results

In this section, we first briefly introduce the experimental setups and then demonstrate the performance of our proposed model of GATMDA by comparing it with seven state-of-the-art methods on two data sets (i.e. HMDAD and Disbiome) under three different cross-validation settings. Finally, we implement case studies on two common diseases to confirm the effectiveness of our model.

### Experimental setup

In this work, we carried out standard 5-fold cross-validation (CV) under the following three different settings:

- CVS1 (overall testing): CV on microbe–disease pairs—random known entries in Y (i.e. microbe–disease pairs) are selected for testing.
- CVS2 (horizontal testing for diseases): CV on diseases—random rows in Y (i.e. diseases) are blinded for testing.
- CVS3 (vertical testing for microbes): CV on microbes—random columns in Y (i.e. microbes) are blinded for testing.

For CVS1, we randomly divide known microbe–disease associations into five groups. For each round, one group of microbe–disease associations (i.e. positive samples) with an equal-size set of unknown randomly sampled microbe–disease pairs (i.e. negative samples) are selected as test samples in turn. And the remaining four groups of microbe–disease associations together with the rest of unknown microbe–disease pairs are utilized to train the model. Specifically, all test samples would first obtain their prediction scores in each round and then be prioritized according to their scores. For a positive (or negative) test sample (microbe–disease pair), we consider that the model successfully predicts the microbe–disease pairs if its ranking is higher (or lower) than a specific threshold. As such, we could obtain the corresponding precision, true positive rates (TPR, sensitivity/recall) and false positive rates (FPR, 1-specificity) by setting different thresholds. Here, precision measures the percentage of the positive test samples among all samples that are predicted as positive with the given threshold. Sensitivity/recall is defined

as the percentage of the positive test samples whose rankings are higher than the given threshold. Specificity means the percentage of the negative test samples that are ranked lower than the given threshold. Subsequently, the receiver operating characteristics (ROCs) curves and precision-recall (PR) curves could be drawn by plotting TPR versus FPR and precision versus recall at different thresholds, respectively. The performance is measured by area under ROC curve (AUC) and area under PR curve (AUPR). To circumvent the influence of random division, each experiment is repeatedly conducted for 10 times. And the final AUC and AUPR scores are calculated over the average of 10 repetitions. Similarly, for CVS2 and CVS3, we randomly sample 20% rows and columns in the adjacent matrix $Y$ as test samples, while the rest of the rows and columns are considered as training samples, respectively. It should be noted that CVS2 and CVS3 are set to predict novel microbe–disease associations for new diseases and new microbes, respectively.

In our model, we set the number of talking-heads $K$ as 4. The number of neural units $r$ was set to 8 in the neural representation aggregation layer. In the procedure of optimization, we limited the influence of weigh matrices with delay factor $\gamma = 0.0005$. While these are our default settings, their influences on the performance of our model will further be discussed in the next section of parameter sensitivity analysis. In addition, the training epoch was set to 200, and the learning rate in the optimization algorithm was set to 0.001. We empirically set the number of hidden units for each attention/head as 4. The dropout was set to 0.5 for HMDAD while 0.3 for Disbiome. The experimental code was implemented based on the open-source machine learning framework Tensorflow (https://github.com/tensorflow/tensorflow). All experiments were conducted on Windows 10 operating system with a HP Z4 G4 workstation computer of an Intel W-2133 8 cores, 3.6GHz CPU and 32G memory.

## Comparison with state-of-the-art methods

To evaluate the performance of our proposed model, we compare GATMDA with seven state-of-the-art methods that were developed for microbe–disease association prediction, including network-based methods, random walk-based methods and matrix factorization-based method.

- KATAHMDA [10] is a KATZ-based computational method.
- NTSHMDA [14] is a random walk with restart-based model.
- BiRWHMDA [13] is a bi-random walk-based model.
- NGRHDMA [39] is a recommendation-based method, which combines neighbour-based collaborative filtering with a graph-based scoring method.
- BRWMDA [15] is a random walk-based method.
- WMGHDMA [12] is a meta-graph-based computational method.
- GRNMFHMDA [17] is a matrix factorization-based model.

For a fair comparison, we ran seven baseline methods on HMDAD data set with their default parameters. For data set Disbiome, we treat the similarity matrices in our model as the input features of diseases and microbes for all baseline methods. Table 2 records the results of 5-fold CV under the setting CVS1 on HMDAD data set. We can observe that among all the methods, our proposed model of GATMDA achieves the best performance with an average AUC of 0.9554±0.0184 and an average AUPR of 0.9334±0.0417, which are 6.37% and 2.72% better than the second-best method NTSHMDA, indicating that our model is effective in predicting novel microbe–disease associations. For better visual comparison, the corresponding ROC and PR curves are drawn in Supplementary Figure S1b and e. In the proposed framework, GATMDA introduces an optimized graph attention network with talking-heads to ensure more informative node representation. Besides, a bi-interaction aggregator is designed to strengthen representation aggregation between similar nodes. In addition, the IMC-based decoder is also an important component to improve prediction accuracy. All of them above contribute to the superior performance of our model, which will be shown in the model ablation in the next section.

In addition, there are several possible reasons why the performance of baseline methods are sub-optimal. First, most of microbes (or diseases) have only one or fewer related diseases (or microbes) in HMDAD. Therefore, only utilizing observed associations may be not enough to identify new associations (e.g. KATZHMDA, NTSHMDA and BiRWHMDA). Second, since the similarity in the network includes some noises, traditional label-propagation algorithms, such as NGRHMDA, BRWMDA, and WMGHMDA, may be limited for this task. Third, the matrix factorization technique can only capture linear associations (e.g. GRNMFHMDA), which is not adequate for the microbe–disease association task.

To further evaluate the effectiveness of our model, we also compare our method with seven baseline methods on the second data set Disbiome. The results in 5-fold CV have been shown in Table 3, which confirms that our model consistently outperforms baseline methods with average AUC of 0.9307±0.0079 and average AUPR of 0.9211±0.0088. Note that all methods show worse on Disbiome than HMDAD. The main reason may be that Disbiome is sparser than HMDAD since the latter's density is 3.95% while the former is only 1.90%. Thus, a model can achieve relatively better training on HMDAD than Disbiome. While all the above results are based on 5-fold CV, we also report the results of various methods using 2-fold CV and 10-fold CV in Supplementary Tables S1 and S2 for HMDAD and Disbiome, respectively. The corresponding ROC and PR curves are also presented in Supplementary Figures S1 and S2, respectively. These results once again verify the effectiveness of our model.

For CVS2 and CVS3, it could be significantly discovered from Table 2 that our method consistently outperforms other baseline methods in terms of AUC and AUPR. Note that the results of some methods are missing under these two scenarios. It is because that these methods depend on known microbe–disease associations for the similarity calculation of diseases and microbes, which makes them unable to achieve prediction when it involves new diseases or new microbes due to the absence of training data. A more direct comparison of different methods under the settings CVS2 and CVS3 could be found in Supplementary Figures S3 and S4, respectively. Overall, our method shows better than baseline methods in identifying novel microbe–disease associations under different scenarios.

From Table 2, we observe that the performance of all methods under CVS1 setting is significantly better than that under CVS2 and CVS3 settings. For new diseases and new microbes, we have no known association pairs for them to train the model, which results in lower performance under CVS2 and CVS3. Furthermore, various methods achieve generally better performance under CVS3 than under CVS2. As the number of microbes (292) is much more than that of diseases (39), the microbe similarity matrix (292x292) is thus more informative than the disease similarity matrix (39x39). Therefore, new microbes can obtain more abundant and accurate information from neighbours than new diseases.

**Table 2.** Performance comparison between seven baseline methods and our model under CVS1, CVS2 and CVS3 settings in 5-fold CV on HMDAD data set. The best results are marked in bold and the second best is underlined. '-' indicates that the corresponding method fails to achieve the task

| Methods | CVS1 | | CVS2 | | CVS3 | |
|---|---|---|---|---|---|---|
| | AUC | AUPR | AUC | AUPR | AUC | AUPR |
| KATZHMDA | 0.8703±0.0199 | 0.8807±0.0167 | – | – | – | – |
| NTSHMDA | 0.8982±0.0312 | 0.9087±0.0294 | – | – | – | – |
| BiRWHMDA | 0.8890±0.0194 | 0.8969±0.0146 | – | – | – | – |
| NGRHMDA | 0.8921±0.0327 | 0.9062±0.0268 | – | – | – | – |
| BRWMDA | 0.8916±0.0029 | 0.9064±0.0152 | 0.6245±0.1221 | 0.6165±0.1233 | – | – |
| WMGHDMA | 0.8745±0.0296 | 0.8895±0.0296 | 0.6123±0.0580 | 0.6774±0.0104 | 0.7511±0.0388 | 0.8313±0.0350 |
| GRNMFHMDA | 0.8806±0.0156 | 0.8914±0.0162 | 0.5743±0.1614 | 0.6224±0.1167 | 0.8820±0.0174 | 0.9004±0.0113 |
| GATMDA | **0.9554±0.0184** | **0.9334±0.0417** | **0.8280±0.0732** | **0.7691±0.0911** | **0.9362±0.0052** | **0.9154±0.0157** |

**Table 3.** Performance comparison between seven baseline methods and our model on Disbiome data set in 5-fold CV under CVS1 setting. The best results are marked in bold and the second best is underlined

| Methods | AUC | AUPR |
|---|---|---|
| KATZHMDA | 0.6779±0.0141 | 0.6785±0.0163 |
| NTSHMDA | 0.8294±0.0071 | 0.7881±0.0099 |
| BiRWHMDA | 0.8344±0.0089 | 0.8104±0.0103 |
| NGRHMDA | 0.8313±0.0052 | 0.8202±0.0043 |
| BRWMDA | 0.8266±0.0031 | 0.8031±0.0041 |
| WMGHDMA | 0.7176±0.0076 | 0.7567±0.0062 |
| GRNMFHMDA | 0.8609±0.0047 | 0.8669±0.0060 |
| GATMDA | **0.9307±0.0079** | **0.9211±0.0088** |

## Ablation study

Recall that our proposed model of GATMDA consists of three components, including (i) GAT with talking-heads-based encoder, (ii) neural representation aggregation layer and (iii) IMC-based decoder. We conduct an ablation study to evaluate the impact of each component in 5-fold CV under the setting CVS1 based on Disbiome data set. We derive the variants of our model as follows:

- GATMDA-GCN: it uses GCN aggregator [40] instead of bi-interaction aggregator in Equation 16, which sums two representations up.
- GATMDA-GraphSage: it uses GraphSage aggregator [41] instead of bi-interaction aggregator in Equation 16, which concatenates two representations.
- GATMDA-M: it uses standard multi-head attention mechanism instead of talking-head attention mechanism in the encoder.
- GATMDA-I: it reconstructs microbe–disease associations using the inner product instead of IMC.

Figure 2 has shown the comparative performance between GATMDA and various variants. We can observe that both of these two terms in Equation 16 are essential components of our mode and play approximately equally important roles in the model, as GATMDA-GCN achieves comparable performance with GATMDA-GraphSage in terms of AUC and AUPR. GATMDA-M achieves lower performance than GATMDA, indicating that talking-head attention mechanism helps improve the prediction accuracy of our model. Besides, we can conclude that IMC can enhance GATMDA's prediction capability, as the AUC and AUPR values of GATMDA-I are lower than that of GATMDA.

## Parameter sensitivity and runtime analysis

There are several important parameters that influence the performance of the model, such as the number of talking-heads $K$, the number of neurons in the bi-interaction aggregator $r$ and the delay factor $\gamma$. It should be noted that we evaluate the influences using 5-fold CV for all parameters based on Disbiome data set. The number of talking-heads $K$ plays an important role in our model. We range $K$ from 1 to 10 with a step value of 1. As shown in Figure 3A, it can be observed that a small or a large value of $K$ is not good for the model performance. The model achieves the best performance when $K = 4$. $r$ is associated with the representation dimension of node. We evaluate the performance of the model by varying $r$ from the range of $\{4, 8, 16, 32, 64, 128, 256, 512\}$. As $r$ increases, the performance first increases and then slightly decreases, with $r = 8$ reaching the best results, as shown in Figure 3B. In addition, we use delay factor $\gamma$ to control the contribution of regularization terms in Equation 18. In our experiment, we vary $\gamma$ from 0.000005 to 0.5 with a step value of 10. From Figure 3C, we can conclude that this parameter has a relatively slight influence on the performance, indicating that our model is robust against the delay factor $\gamma$. The best performance is achieved when $\gamma = 0.0005$.

To evaluate the time complex of our model, we conduct our model on two data sets (i.e. HDMAD and Disbiome) with different sizes to evaluate its running time, respectively. In the proposed framework, GATMDA combines graph attention networks with multi-layer perceptron (MLP) as an encoder to learn representations for microbes and diseases. Here we also compare the running times of GAT (GATMDA-GAT) and MLP (GATMDA-MLP) on different data sets. For each data set, we take the full microbe–disease associations as training data (i.e. 450 association pairs for HMDAD while 2470 for Disbiome). As shown in Figure 4, we observe that as the input data increases exponentially from HMDAD to Disbiome, the running time of GATMDA only increases by 1.17s (the runtime is 0.13s for HMDAD while 1.30s for Disbiome). In addition, we can conclude that MLP takes more time than GAT.

## Case study

To further validate the prediction performance of GATMDA, two common diseases, i.e. asthma and IBD are selected for case studies. For each of them, all known entries are reset to unknown and all candidate microbes are prioritized according to their scores. We carry out our model on HMDAD data set and evaluate the performance by verifying the top 10, 20 and 50 predicted microbes using previous publications.

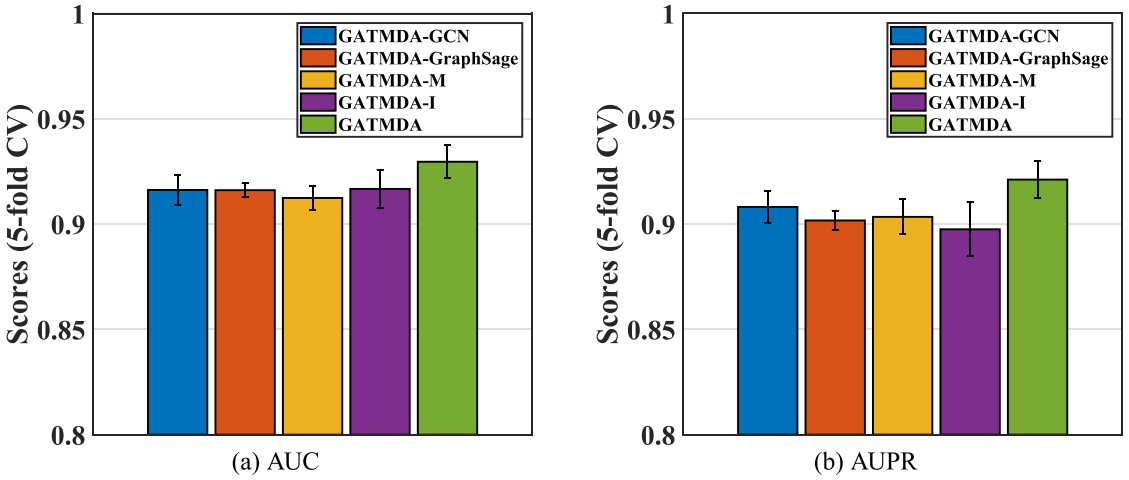


**Figure 2**. Comparative analysis between GATMDA and its variants.

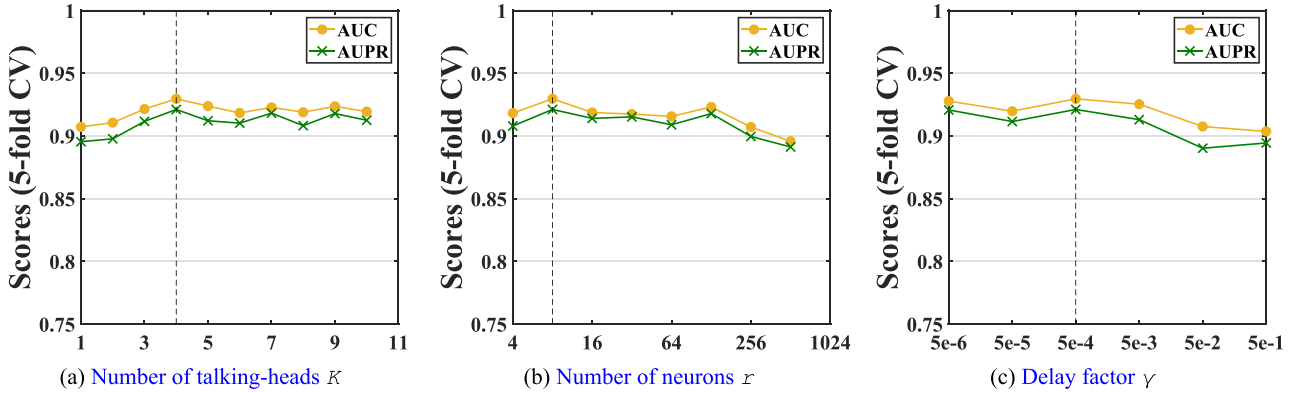


**Figure 3**. Parameter sensitivity under CVS1 for (**A**) number of talking-heads *K*, (**B**) number of neurons *r* and (**C**) delay factor $\gamma$.

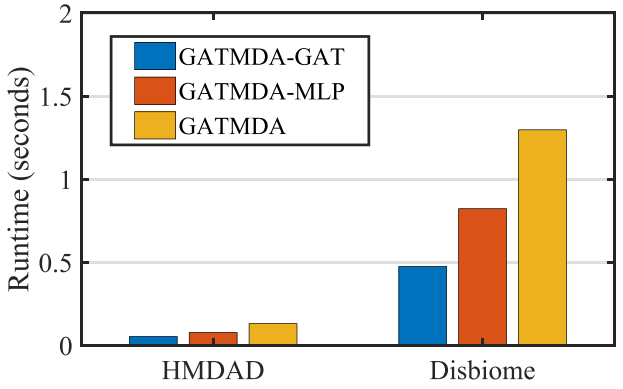


**Figure 4**. Running times of our model on different data sets.

In particular, disease asthma is a common long-term inflammatory disease of the airways of the lungs, which is considered to be caused by both genetic and environmental factors [42]. Accumulated literature has been reported that the microorganisms living in human bodies get involved in the formation and development of asthma. For example, Vael *et al.* [43] demonstrated that *Clostridium coccoides* was significantly associated with asthma, which is predicted by our model to be the most possible candidate microbe for asthma. Marri *et al.* [44] found that *Firmicutes* and *Actinobacteria* were more frequent in samples from nonasthmatic subjects. *Enterobacteriaceae* was shown to be higher in severe asthmatics in comparison with non-severe asthmatics at the family level [45]. As a result, among the top 10, 20 and 50 predicted asthma-related microbes, 10, 18 and 38 disease–microbe associations are confirmed by previous publications. The high prediction accuracy rates, i.e. 100%, 90% and 76%, demonstrate that our model could be used in real-life applications. Table 4 reports the top 50 asthma-associated candidate microbes.

In addition, IBD is a common group of inflammatory conditions of the colon and small intestine [46]. Recent researches have shown that a wide range of microbes are closely associated with IBD. For example, Lloyd *et al.* [47] demonstrated that the activity of IBD can result in molecular disruptions in the transcription among *clostridia*. Walters *et al.* [48] indicated that the abundance of *Bacteroidetes* and *Firmicutes* were significantly increased in IBD subjects compared with healthy controls. Sokol *et al.* [49] uncovered that *Clostridium leptum* and *Clostridium coccoides* groups were less represented in patients with IBD, and *Faecalibacterium prausnitzii* had lower counts in IBD patients. As shown in Table 5, it could be observed that 10, 20 and 44 out of the top 10, 20 and 50 predicted IBD-associated microbes can attain validations from existing reports, indicating that GATMDA has powerful capability in predicting candidate microbes for given diseases and thus is greatly useful for assisting screen candidate target microbes.

**Table 4.** Prediction results of the top 50 asthma-associated microbes. The second column records top 1–25 associated microbes. The fifth column records top 26–50 associated microbes.

| Rank | Microbes | Evidence | Rank | Microbes | Evidence |
|---|---|---|---|---|---|
| 1 | Clostridium coccoides | PMID: 21477358 | 26 | Faecalibacterium prausnitzii | PMID:27253486 |
| 2 | Actinobacteria | PMID:28947029 | 27 | Klebsiella | PMID:29788027 |
| 3 | Firmicutes | PMID:23265859 | 28 | Fusobacterium | PMID: 28486933 |
| 4 | Clostridia | PMID:21477358 | 29 | Burkholderia | PMID:24451910 |
| 5 | Lactobacillus | PMID:20592920 | 30 | Streptococcus | PMID:17950502 |
| 6 | Bifidobacterium | PMID:24735374 | 31 | Porphyromonadaceae | PMID:27433177 |
| 7 | Bacteroides | PMID:18822123 | 32 | Bacteroides uniformis | Uncofirmed |
| 8 | Lachnospiraceae | PMID: 31958431 | 33 | Bacteroidales | PMID: 26870828 |
| 9 | Fusobacterium nucleatum | PMID: 28486933 | 34 | Enterococcus faecium | Uncofirmed |
| 10 | Enterococcus | PMID:29788027 | 35 | Erysipelotrichales | Uncofirmed |
| 11 | Clostridium leptum | PMID:29445257 | 36 | Clostridium | PMID:21477358 |
| 12 | Pseudomonas | PMID: 27076584 | 37 | Bacilli | PMID:28226280 |
| 13 | Veillonella | PMID:25329665 | 38 | Prevotella copri | Uncofirmed |
| 14 | Propionibacterium | PMID: 13268970 | 39 | Oxalobacter formigenes | Uncofirmed |
| 15 | Propionibacterium acnes | PMID: 30185226 | 40 | Bacteroidaceae | PMID:28947029 |
| 16 | Desulfovibrio | PMID: 29198875 | 41 | Enterobacter aerogenes | PMID: 23842440 |
| 17 | Escherichia coli | PMID:29161804 | 42 | Enterobacter hormaechei | Uncofirmed |
| 18 | Enterobacteriaceae | PMID:28947029 | 43 | Klebsiella pneumoniae | PMID: 26953325 |
| 19 | Bacteroides vulgatus | Uncofirmed | 44 | Shigella dysenteriae | Uncofirmed |
| 20 | Verrucomicrobiaceae | Uncofirmed | 45 | Tropheryma whipplei | PMID: 26647445 |
| 21 | Actinomyces | PMID: 32161195 | 46 | Gammaproteobacteria | PMID:28947029 |
| 22 | Porphyromonas gingivalis | PMID: 31342509 | 47 | Betaproteobacteria | Uncofirmed |
| 23 | Selenomonas | PMID: 27093794 | 48 | Clostridium difficile | PMID:21872915 |
| 24 | Treponema | PMID: 31342509 | 49 | Shuttleworthia | Uncofirmed |
| 25 | Bacteroides ovatus | Uncofirmed | 50 | Tannerella | PMID: 31342509 |

## Discussion and Conclusion

Identifying microbe-disease associations can not only provide great insight into understanding the complex pathogenic mechanism of human non-infectious diseases but also boost microbe-oriented therapeutics in precision medicine. For example, systematic identification of potential pathological microbes benefits physicians or biologists in clinically or experimentally discriminating biomarkers for diagnosis and therapeutics [50, 51], especially for complex human diseases. Besides, the computational prediction of disease-causing microbes can help pharmacologists or biologists effectively narrow down the scope of compound candidates [52, 53]. This may further guide them to plan experiments and thus reduce costs. Considering that the conventional wet lab methods are time-consuming, labor-intensive and expensive, computational method provides a great complementary and can guide these experiments. However, previous computational models suffer from two main challenges. On the one hand, most of them are unable to capture the nonlinear associations between diseases and microbes. On the other hand, few models can achieve reasonable predictions for new diseases or new microbes.

In this work, we propose a novel deep learning framework, named GATMDA, based on graph attention network and inductive matrix completion for human microbe-disease association prediction. We fully exploit multiply sources of biological data to construct similarity features for diseases and microbes. To obtain more informative representations, we propose an optimized graph attention network with talking-heads to learn representations for diseases and microbes, which constructs the dependency between different heads and thus enables the model to preserve the importance of high-order neighbours. Besides, for each head, we further design a bi-interaction aggregator in the neural representation aggregation layer to enforce representation aggregation of similar nodes, leading to more accurate node presentations. In addition, we combine the IMC technique to reconstruct disease-microbe associations, which endows the model with the ability to capture the complicated associations between diseases and microbes.

**Table 5.** Prediction results of the top 50 IBD-associated microbes. The second column records top 1–25 associated microbes. The fifth column records top 26–50 associated microbes

| Rank | Microbes | Evidence | Rank | Microbes | Evidence |
|---|---|---|---|---|---|
| 1 | Clostridia | PMID: 31142855 | 26 | Propionibacterium | PMID: 19847949 |
| 2 | Bifidobacterium | PMID:24478468 | 27 | Propionibacterium acnes | PMID: 28630242 |
| 3 | Prevotella | PMID:25307765 | 28 | Rikenellaceae | PMID: 31708890 |
| 4 | Clostridium coccoides | PMID:19235886 | 29 | Ruminococcaceae | PMID: 31379797 |
| 5 | Bacteroides | PMID:25307765 | 30 | Actinomyces | PMID: 30545401 |
| 6 | Firmicutes | PMID:25307765 | 31 | Porphyromonas gingivalis | PMID: 31652577 |
| 7 | Helicobacter pylori | PMID:22221289 | 32 | Selenomonas | Unconfirmed |
| 8 | Bacteroidetes | PMID:25307765 | 33 | Treponema | PMID: 31851086 |
| 9 | Klebsiella | PMID:29573336 | 34 | Desulfovibrio | PMID: 30835854 |
| 10 | Veillonella | PMID:28842640 | 35 | Enterobacteriaceae | PMID:24629344 |
| 11 | Bacteroides ovatus | PMID:30666959 | 36 | Bacteroidaceae | PMID:17897884 |
| 12 | Haemophilus | PMID:24013298 | 37 | Citrobacter | PMID: 30342282 |
| 13 | Fusobacterium | PMID:25307765 | 38 | Pseudomonas | PMID: 31662859 |
| 14 | Staphylococcus | PMID: 27239107 | 39 | Alistipes | PMID:28877044 |
| 15 | Enterococcus | PMID:24629344 | 40 | Parabacteroides | PMID:25307765 |
| 16 | Streptococcus | PMID:23679203 | 41 | Comamonadaceae | Unconfirmed |
| 17 | Bacteroides vulgatus | PMID:29454108 | 42 | Oxalobacteraceae | Unconfirmed |
| 18 | Clostridium leptum | PMID:28099495 | 43 | Sphingomonadaceae | PMID: 27418066 |
| 19 | Fusobacterium nucleatum | PMID: 26718210 | 44 | Enterobacter aerogenes | PMID: 4061480 |
| 20 | Lactobacillus | PMID:26340825 | 45 | Enterobacter hormaechei | Unconfirmed |
| 21 | Bacteroides fragilis | PMID: 31988590 | 46 | Klebsiella pneumoniae | PMID: 9930068 |
| 22 | Bacteroides uniformis | PMID:26789999 | 47 | Shigella dysenteriae | Unconfirmed |
| 23 | Verrucomicrobiaceae | PMID: 22572638 | 48 | Escherichia coli | PMID:29573336 |
| 24 | Porphyromonadaceae | PMID:29573237 | 49 | Faecalibacterium prausnitzii | PMID:24799893 |
| 25 | Porphyromonas | PMID: 31293117 | 50 | Shuttleworthia | Unconfirmed |

Comprehensive experiments demonstrate that the proposed GATMDA model is reliable and promising in identifying potential target microbes for diseases, including both new diseases and new microbes.

However, despite the good prediction performance of our model, there are still some limitations that are expected to be further improved in the future. On the one hand, while our proposed model could predict potential disease-associated microbes, it still cannot determine how the microbial abundances influence disease status. We could further handle this problem by incorporating the microbial abundance information into the network. On the other hand, our model cannot be applied to all new diseases and new microbes, as we fail to obtain the features for new diseases that have no known associated genes, and new microbes that are lack of protein–protein interaction information. In the future, we can collect more prior biological knowledge, such as microbial gene sequencing [54], disease symptom-based similarity [55] and disease semantic similarity [56], to overcome this limitation.

**Key Points**

- *In silico* identification of potential targeted microbes is critical for precision medicine.
- Accumulated available biomedical data provides a golden opportunity to leverage graph-based deep learning techniques to predict novel microbe-disease associations.
- A novel deep learning framework of graph attention networks is developed to infer microbe-disease pairs with IMC.
- Comprehensive experiments demonstrate that GATMDA consistently outperforms seven state-of-the-art methods on different data sets and case studies further confirm the effectiveness of GATMDA in identifying candidate microbes for diseases.

## Supplementary information

Supplementary data are available at *Briefings in Bioinformatics* online

## Funding

This work has been supported by the National Natural Science Foundation of China under (grant numbers: 61873089).

## Conflict of interest

None declared.

## References


1. Cénit M, Matzaraki V, Tigchelaar E, *et al*. Rapidly expanding knowledge on the role of the gut microbiome in health and disease. *Biochim Biophys Acta* 2014; **1842**:1981–92.
2. Huttenhower C, Gevers D, Knight R, *et al*. Structure, function and diversity of the healthy human microbiome. *Nature* 2012; **486**:207.
3. Sommer F, Bäckhed F. The gut microbiota-masters of host development and physiology. *Nat Rev Microbiol* 2013; **11**:227–38.
4. Holmes E, Wijeyesekera A, Taylor-Robinson SD, *et al*. The promise of metabolic phenotyping in gastroenterology and hepatology. *Nat Rev Gastroenterol Hepatol* 2015; **12**:458.
5. Gill SR, Pop M, DeBoy RT, *et al*. Metagenomic analysis of the human distal gut microbiome. *Science* 2006; **312**:1355–9.
6. Henao-Mejia J, Elinav E, Thaiss CA, *et al*. Role of the intestinal microbiome in liver disease. *J Autoimmun* 2013; **46**:66–73.
7. Wen L, Ley RE, Volchkov PY, *et al*. Innate immunity and intestinal microbiota in the development of type 1 diabetes. *Nature* 2008; **455**:1109–13.
8. Huang YJ, Boushey HA. The microbiome in asthma. *J Allergy Clin Immunol* 2015; **135**:25–30.
9. Schwabe RF, Jobin C. The microbiome and cancer. *Nat Rev Cancer* 2013; **13**:800–12.
10. Chen X, Huang Y-A, You Z-H, *et al*. A novel approach based on KATZ measure to predict associations of human microbiota with non-infectious diseases. *Bioinformatics* 2017; **33**:733–9.
11. Huang Z-A, Chen X, Zhu Z, *et al*. PBHMDA: path-based human microbe-disease association prediction. *Front Microbiol* 2017; **8**:233.
12. Long Y, Luo J. WMGHMDA: a novel weighted meta-graph-based model for predicting human microbe-disease association on heterogeneous information network. *BMC Bioinformatics* 2019; **20**:541.
13. Zou S, Zhang J, Zhang Z. A novel approach for predicting microbe-disease associations by bi-random walk on the heterogeneous network. *PLoS One* 2017;**12**:1–16.
14. Luo J, Long Y. NTSHMDA: prediction of human microbe-disease association based on random walk by integrating network topological similarity. *IEEE/ACM Trans Comput Biol Bioinform* 2018. doi: 10.1109/TCBB.2018.2883041.
15. Yan C, Duan G, Wu F, *et al*. Brwmda: predicting microbe-disease associations based on similarities and bi-random walk on disease and microbe networks. *IEEE/ACM Trans Comput Biol Bioinform* 2019. doi: 10.1109/TCBB.2019.2907626.
16. Shen Z, Jiang Z, Bao W. *CMFHMDA: collaborative matrix factorization for human microbe-disease association prediction*, *International Conference on Intelligent Computing*. Liverpool, United Kingdom: Springer, 2017, 261–9.
17. He B-S, Peng L-H, Li Z. Human microbe-disease association prediction with graph regularized non-negative matrix factorization. *Front Microbiol* 2018; **9**:2560.
18. Duan G, Yan C, Wu F, *et al*. Mchmda: predicting microbe-disease associations based on similarities and low-rank matrix completion. *IEEE/ACM Trans Comput Biol Bioinform* 2019. doi: 10.1109/TCBB.2019.2926716.
19. Ma W, Zhang L, Zeng P, *et al*. An analysis of human microbe-disease associations. *Brief Bioinform* 2017; **18**:85–97.
20. Szklarczyk D, Gable AL, Lyon D, *et al*. STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res* 2019; **47**: D607–13.
21. Hwang S, Kim CY, Yang S, *et al*. HumanNet v2: human gene networks for disease research. *Nucleic Acids Res* 2019; **47**:D573–80.
22. Yao L, Mao C, Luo Y. Graph convolutional networks for text classification. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Hawaii, U.S., AAAI Press, 2019, 7370–7.
23. Liu Z, Wan M, Guo S, *et al*. BasConv: aggregating heterogeneous interactions for basket recommendation with graph convolutional neural network. In: *Proceedings of the 2020 SIAM International Conference on Data Mining*. Cincinnati, Ohio, U.S., SIAM, 2020, 64–72.
24. Cai R, Chen X, Fang Y, *et al*. Dual-dropout graph convolutional network for predicting synthetic lethality in human cancers. *Bioinformatics* 2020. doi: 10.1093/bioinformatics/btaa211.
25. Zhang Y, Qi P, Manning CD. Graph convolution over pruned dependency trees improves relation extraction. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels*. Belgium: Association for Computational Linguistics, 2018, 2205–15.
26. Wang X, Ji HY, Shi C, *et al*. Heterogeneous graph attention network. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Hawaii, U.S.: AAAI Press, 2019, 2022–32.
27. Zhang S, Xie L. Improving attention mechanism in graph neural networks via cardinality preservation. In: *International Joint Conference on Artificial Intelligence*. Yokohama, Japan: Morgan Kaufmann, 2020.
28. Wang X, He XN, Cao YX, *et al*. Kgat: knowledge graph attention network for recommendation. In: *Proceedings of the 25th ACMSIGKDD International Conference on Knowledge Discovery & Data Mining*. Alaska, USA: ACM, 2019, 950–8.
29. Wu QT, Zhang HR, Gao XF, *et al*. Dual graph attention networks for deep latent representation of multifaceted social effects in recommender systems. In: *The World Wide Web Conference*. San Francisco, CA, USA: Springer, 2019, 2091–102.
30. Wang L, Huang YC, Hou YL, *et al*. Graph attention convolution for point cloud semantic segmentation. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Long Beach, United States: IEEE, 2019, 10296–305.
31. Zheng K, You Z-H, Wong L, *et al*. Inferring disease-associated Piwi-interacting RNAs via graph attention networks. *bioRxiv*, 2020.
32. Ravindra NG, Sehanobish A, Pappalardo JL, *et al*. Disease state prediction from single-cell data using graph attention networks. In: *Proceedings of the ACM Conference on Health, Inference, and Learning*. Vancouver, Canada: ACM, 2020, 121–30.
33. Janssens Y, Nielandt J, Bronselaer A, *et al*. Disbiome database: linking the microbiome to disease. *BMC Microbiol* 2018; **18**:50.
34. Kamneva OK. Genome composition and phylogeny of microbes predict their co-occurrence in the environment. *PLoS Comput Biol* 2017; **13**:e1005366.
35. Xu J, Li Y. Discovering disease-genes by topological features in human protein–protein interaction network. *Bioinformatics* 2006; **22**:2800–5.
36. Wei H, Liu B. iCircDA-MF: identification of circRNA-disease associations based on matrix factorization. *Brief Bioinform* 2019. doi: 10.1093/bib/bbz057.
37. Velickovic P, Cucurull G, Casanova A, *et al*. Graph attention networks. In: *the 6th International Conference on Learning Representations*. Vancouver, Canada: Vancouver Convention Center, 2018, 1–12.

38. Kingma DP, Ba J. Adam: a method for stochastic optimization. In: the 3rd International Conference on Learning Representations, Hilton San Diego Resort & Spa, 2015. (The publisher of ICLR is unconfirmed).
39. Huang Y-A, You Z-H, Chen X, *et al*. Prediction of microbe-disease association from the integration of neighbor and graph with collaborative recommendation model. *J Transl Med* 2017; **15**:209.
40. Kipf TN, Welling M. Semi-supervised classification with graph convolutional networks. In: *the 5th International Conference on Learning Representations, Palais des Congrès Neptune*. Toulon, France, 2017, 1–14. (The publisher of ICLR is unconfirmed).
41. Hamilton W, Ying Z, Leskovec J. Inductive representation learning on large graphs. In: *Advances in Neural Information Processing Systems*. Long Beach, CA, USA: MIT Press, 2017, 1024–34.
42. Martinez F. Genes, environments, development and asthma: a reappraisal. *Eur Respir J* 2007; **29**:179–84.
43. Vael C, Vanheirstraeten L, Desager KN, *et al*. Denaturing gradient gel electrophoresis of neonatal intestinal microbiota in relation to the development of asthma. *BMC Microbiol* 2011; **11**:68.
44. Marri PR, Stern DA, Wright AL, *et al*. Asthma-associated differences in microbial composition of induced sputum. *J Allergy Clin Immunol* 2013; **131**:346–52.
45. Li N, Qiu R, Yang Z, *et al*. Sputum microbiota in severe asthma patients: relationship to eosinophilic inflammation. *Respir Med* 2017; **131**:192–8.
46. Baumgart DC, Carding SR. Inflammatory bowel disease: cause and immunobiology. *Lancet* 2007; **369**:1627–40.
47. Lloyd-Price J, Arze C, Ananthakrishnan AN, *et al*. Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. *Nature* 2019; **569**:655–62.
48. Walters WA, Xu Z, Knight R. Meta-analyses of human gut microbes associated with obesity and IBD. *FEBS Lett* 2014; **588**:4223–33.
49. Sokol H, Seksik P, Furet J, *et al*. Low counts of Faecalibacterium prausnitzii in colitis microbiota. *Inflamm Bowel Dis* 2009; **15**:1183–9.
50. Takahashi MK, Tan X, Dy AJ, *et al*. A low-cost paper-based synthetic biology platform for analyzing gut microbiota and host biomarkers. *Nat Commun* 2018; **9**: 1–12.
51. Zhou YL, Xu ZJ, He Y, *et al*. Gut microbiota offers universal biomarkers across ethnicity in inflammatory bowel disease diagnosis and infliximab response prediction. *mSystems* 2018; **3**:1–14.
52. Zhou T, Tan L, Cederquist GY, *et al*. 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* 2017; **21**:274–83.
53. Barrows NJ, Campos RK, Powell ST, *et al*. A screen of FDA-approved drugs for inhibitors of Zika virus infection. *Cell Host Microbe* 2016; **20**:259–70.
54. Uchiyama I, Mihara M, Nishide H, *et al*. MBGD update 2018: microbial genome database based on hierarchical orthology relations covering closely related and distantly related comparisons. *Nucleic Acids Res* 2019; **47**: D382–9.
55. Zhang P, Huang X, Li M. Disease prediction and early intervention system based on symptom similarity analysis. *IEEE Access* 2019; **7**:176484–94.
56. Gao XF, Jiang JP, Duan ZQ, *et al*. A new method to measure the semantic similarity from query phenotypic abnormalities to diseases based on the human phenotype ontology. *BMC Bioinformatics* 2018; **19**:162.