Systems Biology

Predicting Human Microbe-Drug Associations via Graph Convolutional Network with Conditional Random Field

Yahui Long [1,2], Min Wu [3], Chee Keong Kwoh [2], Jiawei Luo [1,*] and Xiaoli Li [3,*]

[1]College of Computer Science and Electronic Engineering, Hunan University, Changsha 410000, China
[2]School of Computer Science and Engineering, Nanyang Technological University, 639798, Singapore and
[3]Institute for Infocomm Research, Agency for Science, Technology and Research (A*STAR), 138632, Singapore.

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Human microbes play critical roles in drug development and precision medicine. How to systematically understand the complex interaction mechanism between human microbes and drugs remains a challenge nowadays. Identifying microbe-drug associations can not only provide great insights into understanding the mechanism, but also boost the development of drug discovery and repurposing. Considering the high cost and risk of biological experiments, the computational approach is an alternative choice. However, at present, few computational approaches have been developed to tackle this task.
Results: In this work, we leveraged rich biological information to construct a heterogeneous network for drugs and microbes, including a microbe similarity network, a drug similarity network, and a microbe-drug interaction network. We then proposed a novel Graph Convolutional Network (**GCN**) based framework for predicting human **M**icrobe-**D**rug **A**ssociations, named GCNMDA. In the hidden layer of GCN, we further exploited the Conditional Random Field (CRF), which can ensure that similar nodes (i.e., microbes or drugs) have similar representations. To more accurately aggregate representations of neighborhoods, an attention mechanism was designed in the CRF layer. Moreover, we performed a random walk with restart (RWR) based scheme on both drug and microbe similarity networks to learn valuable features for drugs and microbes respectively. Experimental results on three different datasets showed that our GCNMDA model consistently achieved better performance than seven state-of-the-art methods. Case studies for three microbes including SARS-CoV-2 and two antimicrobial drugs (i.e., *Ciprofloxacin* and *Moxifloxacin*) further confirmed the effectiveness of GCNMDA in identifying potential microbe-drug associations.
Availability: Python codes and dataset are available at: `https://github.com/longyahui/GCNMDA`.
Contact: luojiawei@hnu.edu.cn and xlli@i2r.a-star.edu.sg
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Microbe or microorganism is a category of microscopic living organisms which could be single-celled or multi-cellular. The accumulated evidence have demonstrated that microbe communities, mainly composed of bacteria, archea, viruses, protozoa, and fungi, have close associations with human hosts (Huttenhower *et al.*, 2012; Sommer and Bäckhed, 2013). Microbes are usually viewed as the "forgotten" organ of human beings due to their functions in providing protection from pathogens, improving metabolic capability, and enhancing immunological system (Ventura *et al.*, 2009). For example, the microbes offer protection against invasion by opportunistic pathogens (Sommer and Bäckhed, 2013), facilitate the metabolism of indigestible polysaccharides, and boost T-cell responses via synthesizing essential vitamins (Kau *et al.*, 2011). Besides, they are an imperative component for the development and differentiation of human's intestinal epithelium and immune system (Sommer and Bäckhed, 2013).

On the other hand, the imbalance or dysbiosis of microbe communities could cause a wide range of human infection diseases (Huttenhower *et al.*, 2012; Sommer and Bäckhed, 2013), such as obesity (Zhang *et al.*, 2009), diabetes (Wen *et al.*, 2008), rheumatoid arthritis (Lynch and Pedersen, 2016), and even cancer (Schwabe and Jobin, 2013). As such, the microbes can thus be considered as targets for personalized medicine (Kashyap *et al.*, 2017). In fact, many microbe-drug interactions have been reported in the literature. For example, the microbial β-glucuronidases in the gut assisted the treatment of irinotecan for colorectal cancer by reactivating the excreted, inactive metabolite (Guthrie *et al.*, 2017), and it was found to be an effective inhibitor to reduce CPT-11 induced toxicity (Wallace *et al.*, 2010). Hence, detecting microbe-drug interactions would be very useful for microbe-based therapeutics and drug discovery. However, conventional wet-lab experiments (e.g., culture-based methods) for uncovering microbe-drug associations are time-consuming, laborious, and expensive. Computational approaches for efficiently and accurately predicting microbe-drug associations are thus useful complements to the limited experimental methods.

Recently, several databases are publicly available for *experimentally verified microbe-drug associations*, such as MDAD (Sun *et al.*, 2018), aBiofilm (Rajput *et al.*, 2018) and DrugVirus (Andersen *et al.*, 2020), which enable machine learning techniques to predict *novel* microbe-drug associations. In particular, graph convolutional network (GCN) is a promising machine learning approach due to its superior capability of modeling graph data, which has been successfully used for predicting miRNA-drug resistance association (Huang *et al.*, 2019), disease-gene association (Han *et al.*, 2019) and lncRNA-disease association (Xuan *et al.*, 2019). We were thus motivated to customize GCN for novel microbe-drug association prediction.

Nevertheless, there are two main limitations to the existing GCN based approaches. Firstly, most of them are implemented on either a bipartite network or a homogeneous network to deal with the relevant tasks. Compared with these networks, a *heterogeneous network* can include different types of nodes and links, and it is thus able to leverage diverse and rich semantic information, allowing GCN to better preserve intrinsic features for nodes. Secondly, the graph data possess similarity information between different nodes. However, existing GCN based methods consider all the neighbors equally and thus fail to preserve this kind of similarity information when learning the node embeddings/representations.

To address the above issues, we developed a GCN based framework called GCNMDA for microbe-drug association prediction in a heterogeneous network. First, GCNMDA exploited drug chemical information, microbe gene information and Gaussian interaction profile features to quantify the similarities for drugs and microbes respectively. Considering the noise in similarities, a random walk with restart (RWR) based pre-processing scheme was designed on drug similarity network and microbe similarity network to effectively capture valuable features for drugs and microbes, respectively. Second, we embedded a Conditional Random Field (CRF) layer in GCN to strengthen node representation learning for drugs and microbes, such that similar nodes have similar representations. We further designed an attention mechanism in the CRF layer for more accurately aggregating representations of neighborhoods. Experimental results demonstrated that our proposed GCNMDA model outperformed existing state-of-the-art methods. Case studies on three microbes (i.e., SARS-CoV-2, Pseudomonas aeruginosa, and Escherichia coli) and two popular antibiotic agents (i.e., Ciprofloxacin and Moxifloxacin) further verified the effectiveness of our proposed model.

Overall, our main contributions are summarized as follows.

- We constructed a *heterogeneous network* to effectively integrate rich biological information, including microbe gene information, drug chemical information, and microbe-drug interactions.
- We proposed a novel GCN-based framework for predicting microbe-drug associations in the heterogeneous network. To the best of our knowledge, this is the first work to adapt GCN for predicting microbe-drug associations.
- A CRF layer was designed in GCN, which could enforce that similar nodes (i.e. drugs and microbes) have similar representations. We further designed an attention mechanism in the CRF layer, which assigned greater weight values to more topological similar neighborhoods to preserve similar information between nodes.
- Our comprehensive experimental results and case studies demonstrated the proposed GCNMDA method outperformed seven state-of-the-art methods on three different datasets.

2 Related Work

In this section, we first present graph convolutional networks (GCN) and their applications in bioinformatics. We then introduce conditional random field (CRF) for modelling the dependency between neighboring nodes in the graph. To our best knowledge, so far few approaches were developed for predicting microbe-drug associations.

2.1 Graph convolutional networks

Graph Convolutional Neural Network (GCN), proposed by Kipf and Welling (2016), is an effective deep learning model for graph data. The basic idea of GCN is to learn node embeddings/representations by implementing convolutional operation on a graph based on the properties of neighborhood nodes. In recent years, GCN has achieved great success in a wide range of tasks, such as node classification (Kipf and Welling, 2016), recommender system (Ying *et al.*, 2018) and relation extraction (Zhang *et al.*, 2018).

Very recently, researchers have developed numerous GCN-based approaches to tackle various bioinformatics tasks. For example, Zitnik *et al.* (2018) used graph convolutional network for predicting polypharmacy side effects based on multimodal data. Huang *et al.* (2019) proposed a graph convolutional network-based end-to-end learning framework of GCMDR to address the problem of miRNA-drug resistance association prediction based on a bipartite network. Han *et al.* (2019) developed a new framework named GCN-MF for the identification of disease-gene associations by incorporating graph convolutional network with matrix factorization. To infer candidate disease-related lncRNA, Xuan *et al.* (2019) first constructed a heterogeneous network combining multiple sources of biomedical information. They further presented a combination framework GCNLDA by aggregating graph convolutional network with convolutional neural network. While the above methods achieved good prediction performance, they did not consider node similarities in the hidden layer during the process of representation learning.

2.2 Conditional random field

Conditional random field (CRF), proposed by Lafferty *et al.* (2001), is a probabilistic graphical model. Generally, CRF is applied to predict labels of the sequential data. Its advantage is to model the pairwise relationship between a given node and its neighborhoods to improve the final prediction.

Recently, the combinations of CRF with different deep learning methods have achieved successful applications in various research fields. For example, Liu *et al.* (2015) proposed a novel approach for image segmentation by aggregating Convolutional Neural Network (CNN) with CRF. In addition, Zheng *et al.* (2015) developed a network called CRF-RNN (Recurrent Neural Network) and combined it with CNN for semantic image segmentation. Besides, Cheng *et al.* (2017) also applied CRF-RNN for semantic mapping. In addition, Gao *et al.* (2019) coupled graph

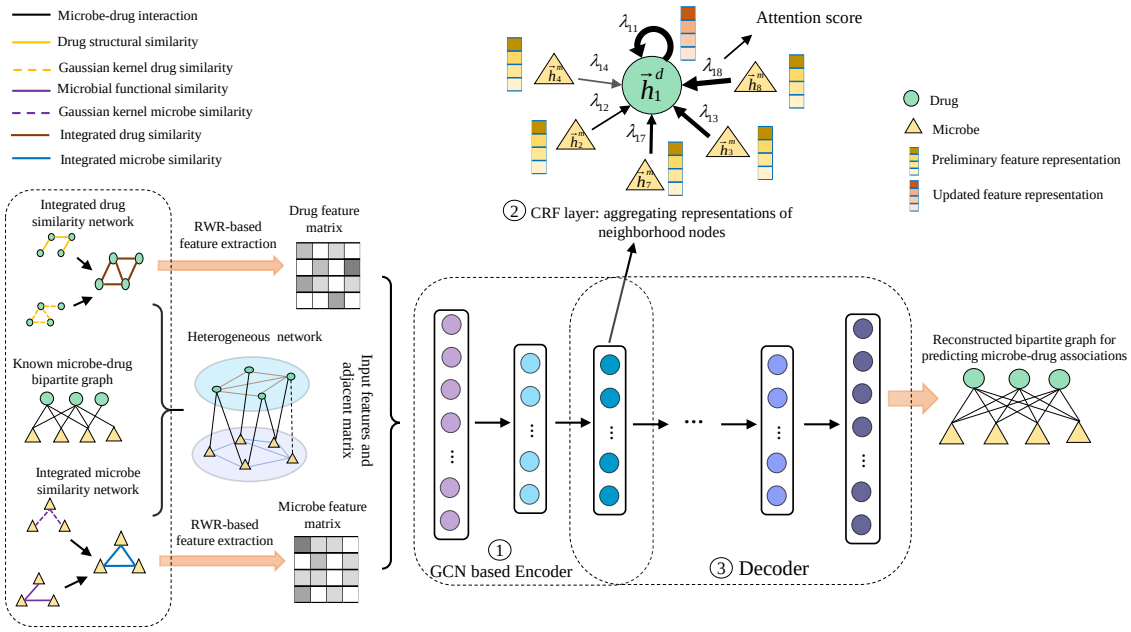


Fig. 1. The overall architecture of GCNMDA for microbe-drug association prediction.

convolutional neural network with CRF for node classification tasks in various homogeneous networks.

3 Materials and Methods

In this work, we propose a novel graph convolutional network (**GCN**) based framework called **GCNMDA** to predict **M**icrobe-**D**rug **A**ssociations. As shown in Figure 1, GCNMDA consists of three main steps. First, we construct *a heterogeneous network* by leveraging rich biological data, including drug similarity, microbe similarity, and microbe-drug bipartite graph. Second, we learn representations for microbes and drugs based on GCN, where a CRF layer is plugged to enforce representation aggregation of neighborhoods. Third, we reconstruct the microbe-drug bipartite network based on the learned representations. Next, we introduce the above three steps in detail.

3.1 Heterogeneous network for microbes and drugs

We use three different datasets for known microbe-drug associations, i.e., MDAD (Sun *et al.*, 2018), aBiofilm (Rajput *et al.*, 2018), and DrugVirus (Andersen *et al.*, 2020). MDAD dataset (`http://www.chengroup.cumt.edu.cn/MDAD/`) consists of 5505 clinically or experimentally verified microbe-drug associations, between 1388 drugs and 174 microbes. After removing redundancy information, we finally obtain 2470 associations between 1373 drugs and 173 microbes. aBiofilm dataset (`http://bioinfo.imtech.res.in/manojk/abiofilm/`) records 1720 unique anti-biofilm agents/drugs which target over 140 organisms/microbes including bacteria and fungus. After filtering out repeated data, we finally download 2884 microbe-drug associations involving 1720 drugs and 140 microbes. DrugVirus dataset (`https://drugvirus.info/tech_doc/`) summarizes activities and developmental statuses of 118 compounds/drugs which altogether target 83 human viruses, including recently occurred novel coronavirus name SARS-CoV-2. Besides, we manually curate 57 clinically or experimentally confirmed drug-virus associations between 76 drugs and 12 viruses from drug databases and related publications. As a result, 933 drug-virus interactions including 175 drugs and 95 viruses are collected. Overall, the statistics of the three microbe-drug association datasets above are shown in Table 1. We define the adjacent matrix $I \in \mathbb{R}^{nd \times nm}$ to represent the microbe-drug associations, where nd and nm denote the numbers of drugs and microbes respectively. I_{ij} is equal to 1 if an association between drug d_i and microbe m_j is observed; 0 otherwise.

Table 1. The statistics for each microbe-drug association dataset.

Datasets	# Microbes	# Drugs	# Associations
MDAD	173	1373	2470
aBiofilm	140	1720	2884
DrugVirus	95	175	933

We further construct *microbe functional similarity matrix* FM and *drug structural similarity matrix* DS. Particularly, FM is calculated by the method proposed by Kamneva (2017). DS is measured using the method SIMCOMP2 (Hattori *et al.*, 2010). More details about the calculations of FM and DS can be found in Supplementary Materials.

It is clear that both FM and DS are sparse, i.e., many microbes or drugs have no similarity scores in FM and DS respectively, due to lacking microbe functional information and drug structural information. To discover more valuable similarity information, we exploit *the Gaussian interaction profile kernel function* to calculate Gaussian kernel similarity for microbes and drugs. The key idea is that *similar microbes (drugs) interact with similar drugs (microbes), leading to similar interaction profiles*. More specifically, in the microbe-drug association matrix I, we define the i-th row $I(d_i)$ and the j-th column $I(m_j)$ as interaction profiles for drug d_i and microbe m_j, respectively. Then, the Gaussian interaction profile kernel similarity matrices GD and GM for drugs and microbes are calculated as follows:

$$GD(d_i, d_j) = \exp(-\eta_d \|I(d_i) - I(d_j)\|^2), \quad (1)$$

$$GM(m_i, m_j) = \exp(-\eta_m \|I(m_i) - I(m_j)\|^2), \quad (2)$$

where η_d and η_m represent the normalized kernel bandwidths, and they are defined in Equations 3 and 4.

$$\eta_d = \eta'_d/(\frac{1}{nd}\sum_{i=1}^{nd}\|I(d_i)\|^2), \tag{3}$$

$$\eta_m = \eta'_m/(\frac{1}{nm}\sum_{i=1}^{nm}\|I(m_i)\|^2), \tag{4}$$

where η'_d and η'_m are the original bandwidths and both are set to 1.

For complementing biological information and improving drug similarity, a final drug similarity is constructed by integrating drug structural similarity and Gaussian kernel drug similarity. Specifically, for drugs d_i and d_j, if there exists drug structural similarity between them, the integrated drug similarity is defined as the average of GD and DS; GD otherwise. The integrated drug similarity S_d is defined as follows:

$$S_d(d_i,d_j) = \begin{cases} \frac{GD(d_i,d_j)+DS(d_i,d_j)}{2}, & if\ DS(d_i,d_j) \neq 0, \\ GD(d_i,d_j), & otherwise. \end{cases} \tag{5}$$

Similarly, the integrated microbe similarity S_m is defined as follows:

$$S_m(m_i,m_j) = \begin{cases} \frac{GM(m_i,m_j)+FM(m_i,m_j)}{2}, & if\ FM(m_i,m_j) \neq 0, \\ GM(m_i,m_j), & otherwise. \end{cases} \tag{6}$$

We finally construct a *heterogeneous network* for drugs and microbes, which consists of three networks: 1) a microbe-drug interaction network, 2) a drug similarity network, and 3) a microbe similarity network. In particular, let $G = (V, E)$ denote the heterogeneous network, with $V = (\nu_m, \nu_d)$ representing a set of nm microbe nodes and nd drug nodes. Its adjacency matrix $A \in \mathbb{R}^{(nd+nm)\times(nd+nm)}$ is defined in Equation 7.

$$A = \begin{bmatrix} S_d & I \\ I^T & S_m \end{bmatrix}. \tag{7}$$

3.2 Feature processing for drugs and microbes

As aforementioned, S_d and S_m are matrices for drug similarity and microbe similarity, respectively. In S_d (or S_m), each row or column denotes the similarity profile for a drug (or a microbe), which can be considered as the feature vector for the drug (or microbe). However, it is insufficient to directly regard the similarity profiles as input features for microbes and drugs because the calculated similarity possibly includes some noises due to false positives and the limitations of computational methods. Therefore, in this paper, we further implement a random walk with restart (RWR) based method to derive the features from the similarity profiles. RWR is a network-based method that can effectively capture local and global topological intrinsic characteristics of a network. Note random walk has been extensively applied for reducing noise in image processing (Jain *et al.*, 2018) and preserving neighbor information in feature learning (Grover and Leskovec, 2016), and thus we adopt it for our problem. Formally, RWR (Köhler *et al.*, 2008) is defined as follows:

$$p_i^{t+1} = (1-\varphi)Mp_i^t + \varphi e_i, \tag{8}$$

where M (i.e., S_d or S_m) represents the transition probability matrix and φ is the restart probability, which is empirically set as 0.9. In addition, $e_i \in \mathbb{R}^{n\times 1}$ is the initial probability vector for the i-th node and e_{ij} is 1 if $j = i$; 0 otherwise. $p_i^t \in \mathbb{R}^{n\times 1}$ shows the probabilities of reaching other nodes at the time t from the i-th node, and we take p_i^t at steady state as the feature vector for the i-th node. After performing RWR on both drug similarity network and microbe similarity network, we obtain a probability profile vector for each microbe or drug. These probability profile vectors can thus form a new drug feature matrix $F_d \in \mathbb{R}^{nd\times nd}$ and a new microbe feature matrix $F_m \in \mathbb{R}^{nm\times nm}$. To make the features comparable among different nodes, we further normalize the probability profile vectors in F_d and F_m, i.e., the sum of probabilities in each vector is normalized to 1. Eventually, the normalized probability profile vectors in F_d and F_m are treated as input features for microbes and drugs in our model. In consistent with the heterogeneous network, a new feature matrix $X \in \mathbb{R}^{(nd+nm)\times(nd+nm)}$ is described as follows:

$$X = \begin{bmatrix} 0 & F_d \\ F_m & 0 \end{bmatrix}. \tag{9}$$

3.3 Graph convolutional network for node embeddings

We have derived the adjacent matrix A in section 3.1 and feature matrix X of heterogeneous network in section 3.2. We can then use them to learn the preliminary embeddings for drugs and microbes using graph convolutional network. The basic idea of GCN is to learn node embeddings by implementing the convolution operation on a graph based on the properties of neighborhood nodes. Formally, let us assume that each node in the heterogeneous network is connected to itself (i.e., self-loop), the normalized systematic adjacent matrix $\widetilde{A}$ of A is defined as $\widetilde{A} = D^{-\frac{1}{2}}AD^{-\frac{1}{2}}$ where D is a diagonal matrix with diagonal elements being $D_{ii} = \sum_{j=1}^{nd+nm} A_{ij}$. Based on these terminologies, the preliminary embeddings $Q \in \mathbb{R}^{(nd+nm)\times n}$ is formulated as follows:

$$Q = \mathrm{ReLU}(\widetilde{A}XW_{en} + B_{en}), \tag{10}$$

where $W_{en} \in \mathbb{R}^{(nd+nm)\times n}$ is a parameter matrix, $B_{en} \in \mathbb{R}^{(nd+nm)\times n}$ is a bias matrix, ReLU (Rectified Linear Unit) represents activation function and n is the dimension of embeddings for drugs and microbes.

3.4 CRF layer for embedding update

After deriving the preliminary embeddings, we further introduce a CRF layer to ensure that similar drugs (or microbes) are also similar in the *feature space*, i.e., have similar embeddings. Meanwhile, we also require a smooth update for the embeddings. As such, we define a loss function $\mathcal{L}_{CRF}$ for this CRF layer in Equation 12, motivated by Gao *et al.* (2019).

$$\mathcal{L}(H_i) = \alpha\|H_i - Q_i\|_2^2 + \beta\sum_{j\in\mathcal{N}_i}\lambda_{ij}\|H_i - H_j\|_2^2, \tag{11}$$

$$L_{CRF} = \sum_{i=1}^{nd+nm}\mathcal{L}(H_i). \tag{12}$$

In Equation 11, Q_i represents the preliminary embedding of node i obtained from the GCN convolution layer and H_i denotes the embedding of node i updated in the CRF layer. In addition, λ denotes attention scores between nodes and λ_{ij} measures the importance of neighbor node j to node i. $\mathcal{N}_i$ is the neighborhood of node i, while α and β are weight factors to balance the influences of the first term and the second term on the prediction performance. The first term in Equation (11) aims to encourage a smooth update for the representation of node i, while the second term in Equation (11) enforces that H_i of node i should be close to H_j of neighbor node j. Meanwhile, we update the node embedding H_i in the CRF layer according to the following rule.

$$H_i^{(k+1)} = \frac{\alpha Q_i + \beta\sum_{j\in\mathcal{N}_i}\lambda_{ij}H_j^{(k)}}{\alpha + \beta\sum_{j\in\mathcal{N}_i}\lambda_{ij}}, \tag{13}$$

where the initial embedding $H_i^{(1)}$ is set as Q_i, and $H_i^{(k)}$ is the embedding updated in the k-th iteration. We set $H_i = H_i^{(K)}$ as the final representation

of node i and K is set as 2 in our experiments. Note that the first layer considers all the neighbours equally, while our proposed CRF layer focuses on similar/important neighbours. In addition, as the number of iteration K in CRF layer increases, the nodes will incrementally gain more and more information from their high-order neighbours.

Moreover, unlike Gao *et al.* (2019), we adopt a self-attention (Vaswani *et al.*, 2017) to differentiate the contributions of neighboring nodes to a given node. Formally, the attention efficient λ_{ij} between node i and node j in Equation 11 is defined as follows.

$$a_{ij} = att(W_t H_i, W_t H_j), \tag{14}$$

$$\lambda_{ij} = softmax(a_{ij}) = \frac{exp(a_{ij})}{\sum_{x \in \mathcal{N}_i} exp(a_{ix})}, \tag{15}$$

where att denotes a single-layer feedforward network to perform the attention, and W_t represents a latent trainable matrix.

3.5 Decoder for microbe-drug association reconstruction

H is the feature/embedding matrix learned in the CRF layer and let us denote the learned feature matrices for drugs and microbes as $H_d \in \mathbb{R}^{nd \times n}$ and $H_m \in \mathbb{R}^{nm \times n}$, respectively. We thus reconstruct the adjacent matrix $Z^{nd \times nm}$ for microbe-drug associations in Equation 16 and derive the reconstruction loss in Equation 17.

$$Z = H_d W_d^{de} (W_m^{de})^T (H_m)^T, \tag{16}$$

$$\mathcal{L}_{REC} = \sum_{(i,j) \in A^+ \cup A^-} \Phi(Z_{ij}, A_{ij}), \tag{17}$$

where $W_d^{de} \in \mathbb{R}^{n \times r}$ and $W_m^{de} \in \mathbb{R}^{n \times r}$ are latent factors that project representations back to original feature space for drugs and microbes, respectively. In addition, Φ is the MSE loss (i.e., mean square error) and A^+ and A^- denote the sets of positive samples and negative samples, respectively.

3.6 Overall loss and optimization

In the encoder and decoder, we have trainable parameters, including W^{en}, B^{en}, W_d^{de} and W_m^{de}. In addition to the losses $\mathcal{L}_{CRF}$ and $\mathcal{L}_{REC}$, we include an regularization term for the model parameters denoted as $\mathcal{L}_\Theta$ in Equation 18. Therefore, the overall loss $\mathcal{L}_{Total}$ is defined in Equation 19.

$$\mathcal{L}_\Theta = \|W^{en}\|^2 + \|B^{en}\|^2 + \left\|W_d^{de}\right\|^2 + \left\|W_m^{de}\right\|^2, \tag{18}$$

$$\mathcal{L}_{Total} = \mathcal{L}_{CRF} + \mathcal{L}_{REC} + \gamma \mathcal{L}_\Theta, \tag{19}$$

where γ is a weight factor.

GCNMDA model is then trained by optimizing the overall loss $\mathcal{L}_{Total}$ above. We employ the Adam optimizer (Kingma and Ba, 2015) for the optimization. Finally, we leverage the scores in the reconstructed matrix Z to rank the unknown pairs for novel microbe-drug association prediction.

4 Results

In this section, we first briefly introduce our experimental setup and then demonstrate the performance of our GCNMDA model by comparing with seven existing methods and the ablation study. Finally, we show the case studies on the top drugs and microbes predicted by our method for three selected microbes and two selected drugs.

4.1 Experimental setup

We conducted three cross-validations (CVs), i.e. 2-fold, 5-fold, and 10-fold, on three different datesets (i.e., MDAD, aBiofilm, and DrugVirus) to evaluate the performance of GCNMDA. Taking 5-fold CV as example, we equally divided all the observed microbe-drug association pairs into 5 groups, and iteratively used 1 group for testing and the remaining 4 groups for training. We reported two well-known performance metrics, widely used for pair-wise association (link) predictions, during cross validation, namely, area under ROC curve (AUC) and area under precision-recall curve (AUPR).

In our model, the training epoch was set to 200 and the learning rate in the optimization algorithm was set to 0.001. In the next section, we will discuss the influences of several other important parameters. 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.

4.2 Comparison with state-of-the-art methods

As mentioned before, few existing approaches have been developed specifically to tackle microbe-drug association prediction problem. Thus, we compare our method with seven *state-of-the-art methods* that were proposed to address other link/association prediction tasks in the field of computational biology.

- KATZHMDA (Chen *et al.*, 2016) is a *KATZ measure based computational method*, developed for microbe-disease prediction.
- WMGHMDA (Long and Luo, 2019) is a *meta-graph-based approach* for predicting microbe-disease associations.
- NTSHMDA (Luo and Long, 2018) is a *random walk with restart based model*, proposed to predict microbe-disease associations.
- IMCMDA (Chen *et al.*, 2018) is a *matrix completion based model* for microRNA-disease association prediction.
- GCMDR (Huang *et al.*, 2019) is a *graph convolutional network based model* for identifying miRNA-drug resistance relationships.
- BLM-NII (Mei *et al.*, 2012) is a *bipartite local model with Neighbor-based Interaction profile Inferring* for drug-target prediction.
- WNN-GIP (Van Laarhoven and Marchiori, 2013) is a *weighted nearest neighbor-Gaussian interaction profile model*, developed to address drug-target prediction problem.

For a fair comparison, all existing seven methods adopted the default parameter values from their original implementations and were compared on the same benchmark MDAD, aBiofilm, and DrugVirus datasets. Table 2 shows the results of 5-fold CV on MDAD. Among all methods, our proposed GCNMDA model achieves the best prediction performance with average AUC of 0.9423±0.0105 and average AUPR of 0.9376±0.0114, which are 2.08% and 1.22% higher than that of the second-best method BLM-NII. It is significantly (at least 7% and 4.54%) better than the remaining six models in terms of AUC and AUPR respectively. To further evaluate the effectiveness and robustness of our model, we also carried out GCNMDA and all the baseline methods on another two datasets aBiofilm and DrugVirus. The results in Table 2 indicate that our model consistently outperforms the baseline methods. In the proposed framework, GCNMDA introduces a CRF layer with attention mechanism in GCN to better aggregate the representations of neighbours. Besides, the RWR-based pre-processing scheme and the decoder are also important components to boost the model performance. They are the main reasons why our proposed model can achieve superior performance, which will be shown in the model ablation study in the next section.

In addition, we also conducted comparative experiments for all methods under 2-fold CV and 10-fold CV settings on these three datasets, as shown in Supplementary Table S1, Table S2, and Table S3. The results

Table 2. Performance comparison between baseline methods and our method on datasets MDAD, aBiofilm and DrugVirus under 5-fold CV. The best results are marked in bold and the second best is underlined.

Methods	MDAD		aBiofilm		DrugVirus	
	AUC	AUPR	AUC	AUPR	AUC	AUPR
KATZHMDA	0.8723±0.0032	0.8384±0.0405	0.9013±0.0072	0.9020±0.0056	0.7809±0.0235	0.7554±0.0314
NTSHMDA	0.8302±0.0089	0.7924±0.0121	0.8213±0.0078	0.7639±0.0095	0.7389±0.0179	0.6973±0.0196
WMGHMDA	0.8654±0.0122	0.8381±0.0083	0.8451±0.0060	0.8903±0.0056	0.7230±0.0214	0.7687±0.0216
IMCMDA	0.7466±0.0102	0.7773±0.0113	0.7750±0.0096	0.8572±0.0049	0.6235±0.0245	0.6962±0.0302
GCMDR	0.8485±0.0062	0.8509±0.0040	0.8772±0.0076	0.8847±0.0061	0.8243±0.0168	0.8206±0.0141
BLM-NII	0.9231±0.0170	0.9263±0.0152	0.9256±0.0842	0.9338±0.0633	0.8913±0.0190	0.8922±0.0221
WNN-GIP	0.8721±0.0162	0.8922±0.0137	0.9019±0.0187	0.9408±0.0132	0.8002±0.0193	0.8436±0.0183
GCNMDA	**0.9423±0.0105**	**0.9376±0.0114**	**0.9517±0.0033**	**0.9488±0.0031**	**0.8986±0.0305**	**0.9038±0.0372**

confirm that GCNMDA once again outperformed all the seven state-of-the-art methods consistently on different datasets, indicating GCNMDA is an effective and robust computational model for predicting microbe-drug associations.

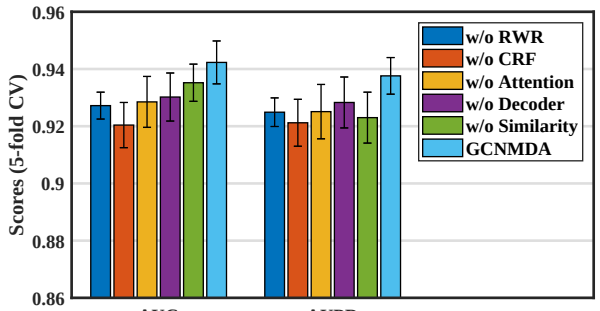


Fig. 2. Comparison analysis between GCNMDA and its Variants on MDAD dataset.

4.3 Model Ablation Study

Recall that our GCNMDA consists of four components, including 1) RWR-based feature processing, 2) CRF layer, 3) attention mechanism in CRF layer, and 4) decoder layer for reconstruction, as shown in Figure 1. In addition, our model is implemented on a heterogeneous network. Here, we conduct the ablation study to evaluate the impact of each component and heterogeneous network using 5-fold CV based on dataset MDAD. In particular, we derive the following model variants for ablation study.

- **GCNMDA w/o RWR**: it uses S_d (or S_m) instead of F_d (or F_m) in Equation 9 as features for drugs (or microbes);
- **GCNMDA w/o CRF**: it has no CRF layer, i.e., $\mathcal{L}_{CRF}$ is not included in the overall loss $\mathcal{L}_{Total}$;
- **GCNMDA w/o Attention**: it uses equal weight instead of bias weight in CRF layer;
- **GCNMDA w/o Decoder**: instead of $H_d W_d^{de} (W_m^{de})^T (H_m)^T$, it uses $H_d (H_m)^T$ in Equation 16 for reconstruction.
- **GCNMDA w/o Similarity**: it uses zero matrices instead of S_d and S_m in Equation 7 and thus its input network is a bipartite network.

Figure 2 shows the performance comparison between GCNMDA and its five variants in terms of AUC and AUPR. We observe that the *CRF layer with attention mechanism* plays the most important role in GCNMDA, as **GCNMDA w/o CRF** achieves the lowest performance. **GCNMDA w/o RWR**, **GCNMDA w/o Attention**, and **GCNMDA w/o Decoder** also achieve lower performance than GCNMDA, indicating that all of RWR-based feature processing, attention mechanism in CRF layer and the decoder are essential components of GCNMDA as they can further enhance GCNMDA's prediction capability. In addition, GCNMDA outperforms **GCNMDA w/o Similarity** in terms of both AUC and AUPR, demonstrating that the heterogeneous network can help our proposed model to achieve better performance than the bipartite network.

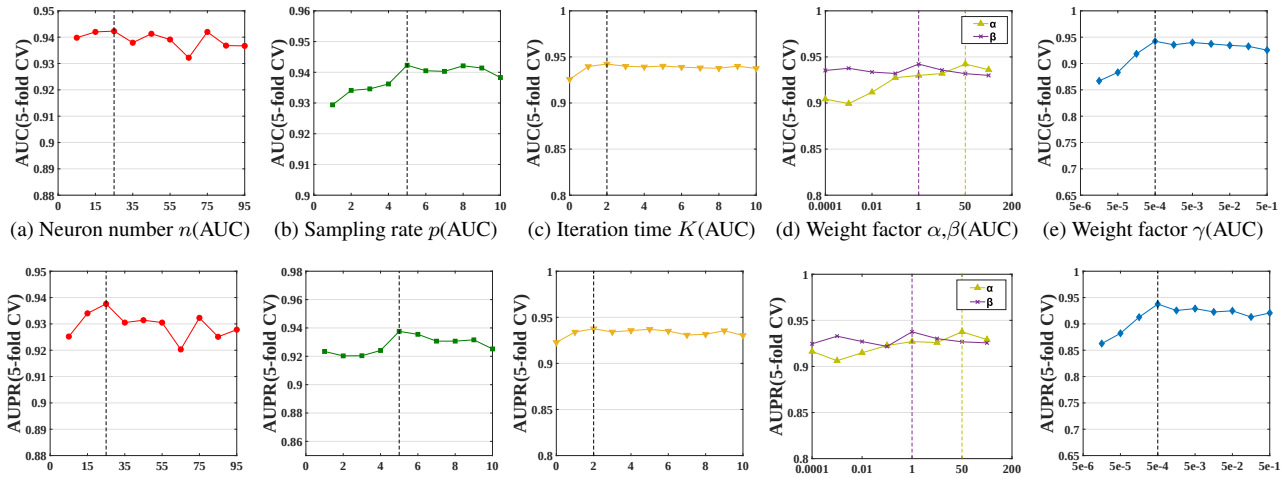
4.4 Parameter sensitivity analysis

In our model, there are several important parameters, such as the neuron number n in hidden layer, the negative sampling rate p, the iteration time K of CRF layer and weight factors α, β and γ. In this section, all experiments were conducted based on dataset MDAD and were evaluated under 5-fold CV. The neuron number may affect the prediction performance of our model. As such, we measure our model performance with different numbers of neurons, ranging from 5 to 95 with a step value of 5. From Figure 3 (a) and (f), we observe that our model is quite robust as both AUC and AUPR values change slightly, i.e. 1-2%, and they reach the best performance with 25 neurons.

As only positive samples exist in the database MDAD, for better model training, we adopt a negative sampling strategy to train our model by randomly selecting some unknown/unlabelled microbe-drug pairs as negative samples. In the experiment, the ratio p, which determines the number of negative samples to that of positive samples, varies from 1 to 10 with a step value of 1. Results in Figure 3 (b) and (g) show, as p increases, its performance *slightly* increases and then decreases, with p=5 achieving its best performance, indicating our model is robust against the parameter p.

In the CRF layer, the iteration number K controls the representation aggregation of high-order neighbors when updating node representations. We evaluate the influence of K by setting its value from 0 to 10 with a step value of 1. Note that $K = 0$ means our model has no CRF layer. It could be found in Figure 3 (c) and (h) that the performance first increases and then slightly decreases. The best performance is reached when $K = 2$. The results indicate that small or large K is not good for the prediction performance of our model.

In the objective function Equation (19), we have weight factors α and β for self-representation and similarity constraint, respectively. Moreover, we employ γ to control the influences of weight matrices on the model. In our experiments, we choose the values from $\{0.0001, 0.001, 0.01, 0.1, 1, 10, 50, 100\}$ for both α and β and the value of γ from $\{0.00001, 0.00005, 0.0001, 0.0005, 0.001, 0.005, 0.01, 0.05, 0.1, 0.5\}$. From Figure 3 (d), (e), (i) and (j), we observe that these parameters have relatively bigger impact on GCNMDA's performance. In particular, GCNMDA achieves better performance when $\alpha = 50$, $\beta = 1$ and $\gamma = 0.0005$, so we set these values as the default ones.



(a) Neuron number n(AUC) (b) Sampling rate p(AUC) (c) Iteration time K(AUC) (d) Weight factor α,β(AUC) (e) Weight factor γ(AUC)

(f) Neuron number n(AUPR) (g) Sampling rate p(AUPR) (h) Iteration time K(AUPR) (i) Weight factor α,β(AUPR) (j) Weight factor γ(AUPR)

Fig. 3. Parameter sensitivity of GCNMDA w.r.t. Neuron number n, Negative sampling rate p, Iteration time K and Weight factors α, β, γ.

Table 3. The top 20 predicted Ciprofloxacin-associated microbes. The first column records top 10 microbes, while the third column records top 11-20 microbes.

Microbe	Evidence	Microbe	Evidence
Candida albicans	PMID:31471074	Enteric bacteria and other eubacteria	PMID:27436461
Streptococcus mutans	PMID:30468214	Listeria monocytogenes	PMID:28355096
Staphylococcus epidermis	PMID:10632381	Burkholderia cenocepacia	PMID:27799222
Staphylococcus epidermidis	PMID:28481197	Streptococcus pneumoniae	PMID:26100702
Enterococcus faecalis	PMID:27790716	Burkholderia pseudomallei	PMID:24502667
Vibrio harveyi	PMID:27247095	Burkholderia multivorans	PMID: 19633000
Salmonella enterica	PMID:26933017	Clostridium perfringens	PMID:29978055
Human immunodeficiency virus 1	PMID:9566552	Serratia marcescens	PMID:23751969
Actinomyces oris	Unconfirmed	Streptococcus epidermidis	Unconfirmed
Streptococcus sanguis	PMID: 11347679	Klebsiella pneumoniae	PMID:27257956

4.5 Case study

To further validate the prediction performance of GCNMDA, we conduct two kinds of case studies based on datasets MDAD and DrugVirus respectively. The first kind of case study includes two popular antimicrobial drugs, i.e. Ciproloxacin and Moxifloxacin, and two popular microbes, i.e. *Pseudomonas aeruginosa* and *Escherichia coli*. For each of them, all the *known* entries are reset to *unknown*, and all candidate microbes (or drugs) are ranked, in decreasing order, according to their predicted scores. We measure the performance of our model by checking whether the most likely candidate microbes (or drugs), occurred in the top 10, 20, and 50 ranking list, are actually verified by previous reports.

Drug Ciprofloxacin is a broad spectrum fluoroquinolone antibacterial agent (Davis *et al.*, 1996), which is mainly applied for therapeutic in most tissues and body fluids. It can show excellent activity against most Gram-negative bacteria. An increasing number of studies have indicated that it has a close interaction with a wide range of human microbes. The common interactions include its activity and toxicity against microorganisms, and the drug-resistance of microorganisms against it. For example, Kim and Woo (2017) indicated that *Enterococcus faecalis* was a kind of high-level ciprofloxacin-resistant bacterial. Zhang *et al.* (2018) found that in vitro experiments, Ciprofloxacin showed a significant killing effectiveness against *Streptococcus mutans*. Hacioglu *et al.* (2019) demonstrated that Ciprofloxacin was an active agent against *Candida albicans*. As a result, among the top 10, 20, and 50 predicted Ciprofloxacin-related microbes, 9, 18, and 41 microbe-drug associations are confirmed by previously published literature, respectively. These high percentages of confirmed microbes, i.e. 90%, 90%, and 82%, demonstrate GCNMDA's capabilities that could be used in real-life applications. Table 3 shows the top 20 predicted candidate microbes associated with Ciprofloxacin. The top 50 predicted candidate microbes can be found in Supplementary Table S4.

Drug Moxiflocacin is a fluoroquinolone antibacterial agent (Balfour and Wiseman, 1999), which has great efficacy in treating patients with respiratory tract, pelvic inflammatory disease (Tulkens *et al.*, 2012) and skin infections (Keating and Scott, 2004). For example, Akiyama and Khan (2012) demonstrated that multiple isolates of *Salmonella enterica* resisted to Moxifloxacin. (Budzinskaya *et al.*, 2019) suggested that in the group of patients with *Pseudomonas aeruginosa* injections, greater resistance to Moxifloxacin was observed compared to the control group. (Dubois and Dubois, 2019) confirmed the bactericidal activity of Moxifloxacin against *Staphylococcus auereus* strains in vitro. By analysing the antibiotic resistance of *Staphylococcus epidermidis*, Eladli *et al.* (2019) found that *Staphylococcus epidermidis* isolated from patient students were susceptible to moxiflocaxin compared to ones isolated from healthy students. Our results show that 9, 16, and 36 out of top 10, 20, and 50 predicted candidate Moxifloxacin-associated microbes are validated by existing reports, indicating GCNMDA has strong capabilities to predict corresponding microbes for given drugs and thus is very useful for drug repositioning. The top 20 and 50 predicted candidate microbes for Moxifloxacin are shown in Table 4 and Supplementary Table S5 respectively.

Pseudomonas aeruginosa is a Gram-negative opportunistic pathogen, which can cause severe acute and chronic infections at different human body sites such as gastrointestinal tracts and skin (Baltch and Smith, 1994). As Supplementary Table S6 shown, 7, 13, 27 out of top 10, 20, and 50 identified candidate *Pseudomonas aeruginosa*-related drugs are

Table 4. The top 20 predicted Moxiflocacin-associated microbes. The first column records top 10 microbes, while the third column records top 11-20 microbes.

Microbe	Evidence	Microbe	Evidence
Pseudomonas aeruginosa	PMID:31691651	Human immunodeficiency virus 1	PMID:18441333
Staphylococcus aureus	PMID:31689174	Actinomyces oris	PMID: 26538502
Escherichia coli	PMID:31542319	Streptococcus sanguis	PMID:10629010
Streptococcus mutans	PMID:29160117	Enteric bacteria and other eubacteria	Unconfirmed
Staphylococcus epidermis	PMID: 11249827	Listeria monocytogenes	PMID:28739228
Staphylococcus epidermidis	PMID:31516359	Burkholderia cenocepacia	Unconfirmed
Enterococcus faecalis	PMID:31763048	Streptococcus pneumoniae	PMID:31542319
Bacillus subtilis	PMID:30036828	Burkholderia pseudomallei	PMID:15731198
Vibrio harveyi	Unconfirmed	Burkholderia multivorans	Unconfirmed
Salmonella enterica	PMID:22151215	Clostridium perfringens	PMID:29486533

Table 5. The top 40 predicted SARS-CoV-2-associated drugs. "**" denotes the predicted drugs with clinical evidences and "*" denotes the predicted drugs with literature supports.

Rank	Drug	Rank	Drug	Rank	Drug	Rank	Drug
1	Favipiravir**	11	Brincidofovir	21	Tilorone (Amixin)	31	Foscarnet
2	Mycophenolic acid*	12	Sorafenib	22	Suramin	32	Azithromycin**
3	Nitazoxanide*	13	Gemcitabine	23	Labyrinthopeptin A2	33	Mitoxantrone
4	Cidofovir*	14	Monensin	24	Luteolin	34	Amiodarone
5	Obatoclax	15	Eflornithine	25	Letermovir	35	Raloxifene
6	Amodiaquine	16	Glycyrrhizin	26	Rapamycin (Sirolimus)	36	Pentosan polysulfate
7	Emetine	17	Berberine	27	Artesunate	37	Bortezomib
8	Niclosamide	18	ABT-263	28	Emodin	38	Labyrinthopeptin A1
9	Brequinar	19	BCX4430 (Galidesivir)	29	Chlorpromazine	39	Silvestrol
10	EIPA (amiloride)	20	Cyclosporine*	30	Ganciclovir	40	Sunitinib

verified by previous literatures. *Escheriachia coli* is a Gram-negative, rod-shaped, coliform bacterium, which is commonly found in human intestine (Tenaillon *et al.*, 2010). Most *Escheriachia coli* are harmless, but some strains can cause disease of the gastrointestinal, urinary, or central nervous system. (Nataro and Kaper, 1998). The results generated by GCNMDA indicate that 8, 14, 22 out of top 10, 20, and 50 predicted candiated *Escheriachia coli*-associated drugs could be confirmed by existing reports, as shown in Supplementary Table S7.

Given that COVID-19 has caused great damages to human life globally, it is thus both interesting and timely to leverage our developed method to address this serious problem. In particular, we performed an additional case study on DrugVirus dataset for SARS-CoV-2, which is the etiologic agent of COVID-19. In particular, we reset all the known SARS-CoV-2-related entries as unknown ones and prioritize the drugs for SARS-CoV-2 according to their prediction scores. At present, no specific antivirals or approved vaccines are available to combat COVID-19 (Hoffmann *et al.*, 2020). However, several drugs such as favipiravir, chloroquine, arbidol, remdesivir, and azithromycin are currently undergoing clinical studies to test their efficacy and safety in the treatment of COVID-19 (Dong *et al.*, 2020). Table 5 shows the top 40 predicted associated drugs for SARS-CoV-2, where some drugs have been successfully verified by previous clinical experiments. For example, Cai *et al.* (2020) demonstrated that favipiravir showed great therapeutic responses to COVID-19 in the clinical experiment, and favipiravir is ranked by our model as the first candidate drug for SARS-CoV-2. Gautret *et al.* (2020) indicated that the combination of azithromycin and hydroxychloroquine could clinically improve the disease of patients with COVID-19, while azithromycin is predicted by our model as the top 32^{nd} drug associated with SARS-CoV-2.

In addition, some predicted drugs have been considered to be possible or referenced treatment drugs for COVID-19. For example, Lee *et al.* (2020) indicated that the patients with COVID-19 could receive treatments of some systematic antibiotics and antiviral medications, such as Mycophenolic acid (the 2^{nd}) and Cyclosporine (the 20^{th}). In addition, Kelleni (2020) discovered that the combination of Nitazoxanide (the 3^{rd}) with Azithromycin (the 32^{nd}) has potential antiviral activity against SARS-COV-2. Through analyzing the structural and chemical features of FDA-approved drugs, Jockusch *et al.* (2020) showed that Cidofovir (the 4^{th}) is likely a potential therapy for COVID-19. The corresponding prediction scores for these 40 candidate drugs can be found in Supplementary Table S8.

5 Discussion and conclusion

Recent research have clearly shown human microbes residing within and upon human bodies play critical roles for human health. Predicting microbe-drug associations can benefit human beings by facilitating the efficient development of drugs and personalized medicine. Compared with conventional culture-based methods, computational methods are able to more effectively identify target microbes for existing drugs or new drugs for known microbes, on a global scale. However, to date, we have found few computational methods to address this important problem, possibly because only recently some experimental validated microbe-drug associations become available for designing computational methods.

In this paper, we present a novel graph convolutional neural network-based framework, named GCNMDA, for predicting new microbe-drug associations. In particular, we first construct a *heterogeneous network* to effectively integrate rich biological information, including microbe gene information, drug chemical information, and microbe-drug interactions. We then implement a RWR based pre-processing mechanism for effective feature extraction. Finally, we introduce an additional CRF layer in GCN, which could enforce that similar nodes (i.e. drugs and microbes) have similar representations. We further design an attention mechanism in the CRF layer, which assigns greater weight values to more similar neighborhoods for preserving topological similar information between nodes, leading to more accurate node representations. Extensive experimental results and case studies demonstrate that the

proposed GCNMDA method significantly outperforms seven state-of-the-art methods in predicting microbe-drug associations on the benchmark MDAD dataset, as well as its great potential for drug discovery.

Although we leverage multiple types of prior biological information to construct similarities for microbes and drugs, there is still room to improve our prediction model through further data integration. In the future, we can incorporate more biological information, such as microbe-disease associations (Chen *et al.*, 2016), microRNA-disease associations (Xiao *et al.*, 2018; Chen *et al.*, 2019), LncRNA-disease associations (Chen *et al.*, 2017) and drug-target associations (Liu *et al.*, 2016; Ezzat *et al.*, 2019), for microbe-drug association prediction. More specifically, we can exploit such information to enrich input features for microbes and drugs or construct multiplex and heterogeneous biological networks (Valdeolivas *et al.*, 2019) to improve the prediction performance of our model.

Acknowledgements

This work has been supported by the National Natural Science Foundation of China (Grant no.61873089) and the Chinese Scholarship Council (CSC) (201906130027).

References

Akiyama, T. and Khan, A. A. (2012). Isolation and characterization of small qnrs1-carrying plasmids from imported seafood isolates of salmonella enterica that are highly similar to plasmids of clinical isolates. *FEMS Immunology & Medical Microbiology*, **64**(3), 429–432.

Andersen, P. I. *et al.* (2020). Discovery and development of safe-in-man broad-spectrum antiviral agents. *International Journal of Infectious Diseases*.

Balfour, J. A. B. and Wiseman, L. R. (1999). Moxifloxacin. *Drugs*, **57**(3), 363–373.

Baltch, A. L. and Smith, R. P. (1994). Pseudomonas aeruginosa: infections and treatment. *Pseudomonas aeruginosa: infections and treatment.*, (12).

Budzinskaya, M. *et al.* (2019). Conjunctival microflora and its antibiotic sensitivity after serial intravitreal injections. *Vestnik oftalmologii*, **135**(5. Vyp. 2), 135–140.

Cai, Q. *et al.* (2020). Experimental treatment with favipiravir for covid-19: an open-label control study. *Engineering*.

Chen, X. *et al.* (2016). A novel approach based on katz measure to predict associations of human microbiota with non-infectious diseases. *Bioinformatics*, **33**(5), 733–739.

Chen, X. *et al.* (2017). Long non-coding rnas and complex diseases: from experimental results to computational models. *Briefings in bioinformatics*, **18**(4), 558–576.

Chen, X. *et al.* (2018). Predicting mirna–disease association based on inductive matrix completion. *Bioinformatics*, **34**(24), 4256–4265.

Chen, X. *et al.* (2019). Micrornas and complex diseases: from experimental results to computational models. *Briefings in bioinformatics*, **20**(2), 515–539.

Cheng, J. *et al.* (2017). A dense semantic mapping system based on crf-rnn network. In *2017 18th International Conference on Advanced Robotics (ICAR)*, pages 589–594. IEEE.

Davis, R. *et al.* (1996). Ciprofloxacin. *Drugs*, **51**(6), 1019–1074.

Dong, L., Hu, S., and Gao, J. (2020). Discovering drugs to treat coronavirus disease 2019 (covid-19). *Drug discoveries & therapeutics*, **14**(1), 58–60.

Dubois, J. and Dubois, M. (2019). Levonadifloxacin (wck 771) exerts potent intracellular activity against staphylococcus aureus in thp-1 monocytes at clinically relevant concentrations. *Journal of medical microbiology*, **68**(12), 1716–1722.

Eladli, M. G. *et al.* (2019). Antibiotic-resistant staphylococcus epidermidis isolated from patients and healthy students comparing with antibiotic-resistant bacteria isolated from pasteurized milk. *Saudi journal of biological sciences*, **26**(6), 1285–1290.

Ezzat, A. *et al.* (2019). Computational prediction of drug–target interactions using chemogenomic approaches: an empirical survey. *Briefings in bioinformatics*, **20**(4), 1337–1357.

Gao, H. *et al.* (2019). Conditional random field enhanced graph convolutional neural networks. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 276–284.

Gautret, P. *et al.* (2020). Clinical and microbiological effect of a combination of hydroxychloroquine and azithromycin in 80 covid-19 patients with at least a six-day follow up: A pilot observational study. *Travel Medicine and Infectious Disease*, page 101663.

Grover, A. and Leskovec, J. (2016). node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 855–864.

Guthrie, L. *et al.* (2017). Human microbiome signatures of differential colorectal cancer drug metabolism. *NPJ biofilms and microbiomes*, **3**(1), 27.

Hacioglu, M. *et al.* (2019). Effects of ceragenins and conventional antimicrobials on candida albicans and staphylococcus aureus mono and multispecies biofilms. *Diagnostic microbiology and infectious disease*, **95**(3), 114863.

Han, P. *et al.* (2019). Gcn-mf: Disease-gene association identification by graph convolutional networks and matrix factorization. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 705–713.

Hattori, M. *et al.* (2010). Simcomp/subcomp: chemical structure search servers for network analyses. *Nucleic acids research*, **38**(suppl_2), W652–W656.

Hoffmann, M. *et al.* (2020). Sars-cov-2 cell entry depends on ace2 and tmprss2 and is blocked by a clinically proven protease inhibitor. *Cell*.

Huang, Y.-a. *et al.* (2019). Graph convolution for predicting associations between mirna and drug resistance. *Bioinformatics*.

Huttenhower, C. *et al.* (2012). Structure, function and diversity of the healthy human microbiome. *nature*, **486**(7402), 207.

Jain, D. K. *et al.* (2018). Random walk-based feature learning for micro-expression recognition. *Pattern Recognition Letters*, **115**, 92–100.

Jockusch, S. *et al.* (2020). A library of nucleotide analogues terminate rna synthesis catalyzed by polymerases of coronaviruses causing sars and covid-19. *bioRxiv*.

Kamneva, O. K. (2017). Genome composition and phylogeny of microbes predict their co-occurrence in the environment. *PLoS computational biology*, **13**(2), e1005366.

Kashyap, P. C. *et al.* (2017). Microbiome at the frontier of personalized medicine. In *Mayo Clinic Proceedings*, pages 1855–1864.

Kau, A. L. *et al.* (2011). Human nutrition, the gut microbiome and the immune system. *Nature*, **474**(7351), 327.

Keating, G. M. and Scott, L. J. (2004). Moxifloxacin. *Drugs*, **64**(20), 2347–2377.

Kelleni, M. (2020). Nitazoxanide/azithromycin combination for covid-19: A suggested new protocol for covid-19 early management. *Pharmacological Research*, **157**, 104874.

Kim, M.-C. and Woo, G.-J. (2017). Characterization of antimicrobial resistance and quinolone resistance factors in high-level ciprofloxacin-resistant enterococcus faecalis and enterococcus faecium isolates obtained from fresh produce and fecal samples of patients. *Journal of the Science of Food and Agriculture*, **97**(9), 2858–2864.

Kingma, D. P. and Ba, J. (2015). Adam: A method for stochastic optimization. *ICLR*.

Kipf, T. N. and Welling, M. (2016). Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.

Köhler, S. *et al.* (2008). Walking the interactome for prioritization of candidate disease genes. *The American Journal of Human Genetics*, **82**(4), 949–958.

Lafferty, J. *et al.* (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the 18th International Conference on Machine Learning*, pages 282–289.

Lee, C.-H. *et al.* (2020). Role of dermatologists in the uprising of the novel corona virus (covid-19): Perspectives and opportunities. *Dermatologica Sinica*, **38**(1), 1.

Liu, F. *et al.* (2015). Crf learning with cnn features for image segmentation. *Pattern Recognition*, **48**(10), 2983–2992.

Liu, Y. *et al.* (2016). Neighborhood regularized logistic matrix factorization for drug-target interaction prediction. *PLoS computational biology*, **12**(2).

Long, Y. and Luo, J. (2019). Wmghmda: a novel weighted meta-graph-based model for predicting human microbe-disease association on heterogeneous information network. *BMC Bioinformatics*, **21**(1), 541.

Luo, J. and Long, Y. (2018). Ntshmda: Prediction of human microbe-disease association based on random walk by integrating network topological similarity. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, pages 1–1.

Lynch, S. V. and Pedersen, O. (2016). The human intestinal microbiome in health and disease. *New England Journal of Medicine*, **375**(24), 2369–2379.

Mei, J.-P. *et al.* (2012). Drug–target interaction prediction by learning from local information and neighbors. *Bioinformatics*, **29**(2), 238–245.

Nataro, J. P. and Kaper, J. B. (1998). Diarrheagenic escherichia coli. *Clinical microbiology reviews*, **11**(1), 142–201.

Rajput, A. *et al.* (2018). abiofilm: a resource of anti-biofilm agents and their potential implications in targeting antibiotic drug resistance. *Nucleic acids research*, **46**(D1), D894–D900.

Schwabe, R. F. and Jobin, C. (2013). The microbiome and cancer. *Nature Reviews Cancer*, **13**(11), 800.

Sommer, F. and Bäckhed, F. (2013). The gut microbiota—masters of host development and physiology. *Nature Reviews Microbiology*, **11**(4), 227.

Sun, Y.-Z. *et al.* (2018). Mdad: a special resource for microbe-drug associations. *Frontiers in cellular and infection microbiology*, **8**.

Tenaillon, O. *et al.* (2010). The population genetics of commensal escherichia coli. *Nature Reviews Microbiology*, **8**(3), 207–217.

Tulkens, P. M. *et al.* (2012). Moxifloxacin safety. *Drugs in R&D*, **12**(2), 71–100.

Valdeolivas, A. *et al.* (2019). Random walk with restart on multiplex and heterogeneous biological networks. *Bioinformatics*, **35**(3), 497–505.

Van Laarhoven, T. and Marchiori, E. (2013). Predicting drug-target interactions for new drug compounds using a weighted nearest neighbor profile. *PloS one*, **8**(6), e66952.

Vaswani, A. *et al.* (2017). Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Ventura, M. *et al.* (2009). Genome-scale analyses of health-promoting bacteria: probiogenomics. *Nature Reviews Microbiology*, **7**(1), 61.

Wallace, B. D. *et al.* (2010). Alleviating cancer drug toxicity by inhibiting a bacterial enzyme. *Science*, **330**(6005), 831–835.

Wen, L. *et al.* (2008). Innate immunity and intestinal microbiota in the development of type 1 diabetes. *Nature*, **455**(7216), 1109.

Xiao, Q. *et al.* (2018). A graph regularized non-negative matrix factorization method for identifying microrna-disease associations. *Bioinformatics*, **34**(2), 239–248.

Xuan, P. *et al.* (2019). Graph convolutional network and convolutional neural network based method for predicting lncrna-disease associations. *Cells*, **8**(9), 1012.

Ying, R. *et al.* (2018). Graph convolutional neural networks for web-scale recommender systems. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 974–983. ACM.

Zhang, H. *et al.* (2009). Human gut microbiota in obesity and after gastric bypass. *Proceedings of the National Academy of Sciences*, **106**(7), 2365–2370.

Zhang, Y. *et al.* (2018). Graph convolution over pruned dependency trees improves relation extraction. *arXiv preprint arXiv:1809.10185*.

Zheng, S. *et al.* (2015). Conditional random fields as recurrent neural networks. In *Proceedings of the IEEE international conference on computer vision*, pages 1529–1537.

Zitnik, M. *et al.* (2018). Modeling polypharmacy side effects with graph convolutional networks. *Bioinformatics*, **34**(13), i457–i466.