BMC Bioinformatics

RESEARCH

Open Access

GACNNMDA: a computational model for predicting potential human microbe-drug associations based on graph attention network and CNN-based classifier

Qing Ma[1], Yaqin Tan[2,3*] and Lei Wang[1,2,3*]

*Correspondence:
1970250317@qq.com;
wanglei@xtu.edu.cn

[1] School of Software and Information Engineering, Hunan Software Vocational and Technical University, Xiangtan 411108, China
[2] Big Data Innovation and Entrepreneurship Education Center of Hunan Province, Changsha University, Changsha 410022, China
[3] Key Laboratory of Hunan Province for Internet of Things and Information Security, Xiangtan University, Xiangtan 411105, China

Abstract

As new drug targets, human microbes are proven to be closely related to human health. Effective computational methods for inferring potential microbe-drug associations can provide a useful complement to conventional experimental methods and will facilitate drug research and development. However, it is still a challenging work to predict potential interactions for new microbes or new drugs, since the number of known microbe-drug associations is very limited at present. In this manuscript, we first constructed two heterogeneous microbe-drug networks based on multiple measures of similarity of microbes and drugs, and known microbe-drug associations or known microbe-disease-drug associations, respectively. And then, we established two feature matrices for microbes and drugs through concatenating various attributes of microbes and drugs. Thereafter, after taking these two feature matrices and two heterogeneous microbe-drug networks as inputs of a two-layer graph attention network, we obtained low dimensional feature representations for microbes and drugs separately. Finally, through integrating low dimensional feature representations with two feature matrices to form the inputs of a convolutional neural network respectively, a novel computational model named GACNNMDA was designed to predict possible scores of microbe-drug pairs. Experimental results show that the predictive performance of GACNNMDA is superior to existing advanced methods. Furthermore, case studies on well-known microbes and drugs demonstrate the effectiveness of GACNNMDA as well. Source codes and supplementary materials are available at: https://github.com/tyqGitHub/TYQ/tree/master/GACNNMDA

Keywords: Microbe-drug associations, Graph attention network, Convolutional neural network, Computational model, Prediction model

Background

Researches show that Microorganisms play an integral and often unique role in human beings [1]. The microbiota and its metabolites are essential to the regulation of the host metabolism and immunity [2]. Microbes have a great impact on human health in many

ways, including resistance to the invasion of opportunistic pathogens [3], promotion of the synthesis of sugar metabolism and synthesis of the necessary vitamins to boost T-cell responses [4], etc. In recent years, different aspects of the microbiome and its potential role in human health, including the early life and specific diseases, have been widely reported. For instance, Sprockett et al. explored how priority effects might influence microbial communities in the gastrointestinal tract during early childhood [5]. Ximenez et al. discussed the development of microbiota during the early times of life, from pregnancy to delivery to the early years after birth [6]. And in addition, it has been demonstrated that the intestinal microbiota plays a key role in cardiometabolic disorders, inflammatory bowel diseases, neuropsychiatric diseases and cancer separately [7–12]. Moreover, bacteria and viruses have been proven to be able to cause infectious diseases such as COVID-19 as well [13].

Simultaneously, studies show that when using drugs to treat diseases, not only the administration of drugs can affect the microbiome, but also microbial metabolism can significantly affect the clinical response of drugs [14, 15]. For example, penicillin is an important antibiotic with high efficiency and has treated pneumonia, meningitis, endocarditis, diphtheria, anthrax and so on. However, the widespread use of antibiotics has led to the development of resistance in human microbes such as staphylococcus aureus and Escherichia coli. As a result, there is an urgent need to uncover potential associations between microbes and drugs for drug development. Considering that traditional bio-experiments are quite expensive and time-consuming, it is meaningful to develop calculation models to infer possible associations between microbes and drugs, because these models can be used to guide the experimental designs of wet-lab experiments efficiently.

With the development of bioinformatical technologies, in recent years, several well-known public microbe-drug association databases such as MDAD [16], aBiofilm [17] and Drugvirus [18] have been constructed successively. Based on these databases, researchers around the world have proposed a large number of prediction methods that can be utilized to identify latent associations between microbe-drug pairs. For example, though introducing the KATZ metric to detect possible associations between microbe-drug pairs, Zhu et al. designed a prediction model named HMDAKATZ [19]. By integrating the metapath2vec scheme with a bipartite network recommendation algorithm, Long et al. proposed a computational approach called HNERMDA to infer microbe-drug associations [20]. Additionally, in 2021, Zhu et al. introduced a novel Laplacian Regularized Least Square based prediction method called LRLSMDA, which can discover latent associations between microbe-drug pairs effectively [21]. In the literature [22], through combining the graph convolutional network (GCN) with the conditional random field (CRF), Long et al. conceived a calculative model named GCNMDA to predict possible microbe-drug associations. In the literature [23], Long et al. constructed a framework of graph attention networks called EGATMDA for latent microbe–drug association prediction. Furthermore, In 2022, Deng et al. designed a multi-modal variational graph embedding model named Graph2MDA for prediction of possible microbe–drug associations [24].

Inspired by above methods, through combining the graph attention network (GAT) with a convolutional neural network (CNN)-based classifier, we proposed a novel computational model called GACNNMDA to discover potential microbe-drug associations in this manuscript. In GACNNMDA, through combining multiple measures of similarity

of microbes and drugs, with known microbe-drug associations or known microbe-disease-drug associations respectively, we constructed two heterogeneous microbe-drug networks first. And then, by leveraging multiple types of microbe and drug features, we established two feature matrices for microbes and drugs simultaneously. Thereafter, after inputting these two feature matrices and two heterogeneous microbe-drug networks into a two-layer graph attention network (GAT), we obtained low dimensional feature representations for microbes and drugs respectively. Finally, we designed a convolutional neural network (CNN)-based classifier to predict possible scores of microbe-drug pairs, by integrating low dimensional feature representations and two feature matrices to form the inputs. Moreover, in order to verify the predictive performance of GACNNMDA, we performed intensive comparison experiments and case studies. Experimental results demonstrated that GACNNMDA outperformed existing representative competitive methods, and can achieve satisfactory performances in latent microbe-drug association prediction.

Data sources

Firstly, we will download known microbe-drug associations from the database MDAD (http://www.chengroup.cumt.edu.cn/MDAD/), which includes 2470 clinically or experimentally verified microbe-drug associations between 1373 drugs and 173 microbes.

Secondly, we will download known associations among microbes, drugs and diseases from the dataset collected by Wang et al. [25], which consists of 70,315 known drug-disease associations and 15,633 known microbe-disease associations. After removing those associations associated with diseases that have no known association with any drug or microbe included in MDAD, we obtained 1121 different drug-disease associations between 233 drugs and 109 diseases, and 402 different microbe-disease associations between 73 microbes and 109 diseases respectively.

Finally, from the dataset constructed by Deng et al. [24], we collected 5586 known drug-drug interactions covering 1228 drugs in MDAD, and 138 microbe-microbe interactions covering 123 microbes in MDAD, separately. Details of these aforementioned data were shown in the following Table 1.

For convenience, all these newly downloaded datasets of diseases, drugs, microbes, druf-disease associations, drug-drug interactions, microbe-drug associations, microbe-disease associations and microbe-microbe interactions will be kept in Additional files 1–8 separately.

Methods

As shown in Fig. 1, GACNNMDA mainly consists of three parts:

Table 1 Details of our downloaded data

Type	Microbes	Drugs	Diseases	Associations
Microbe-drug associations	173	1373	–	2470
Microbe-disease associations	73	–	109	402
Drug-disease associations	–	233	109	1121
Drug–drug interactions	–	1228	–	5586
Microbe–microbe interactions	123	–	–	138

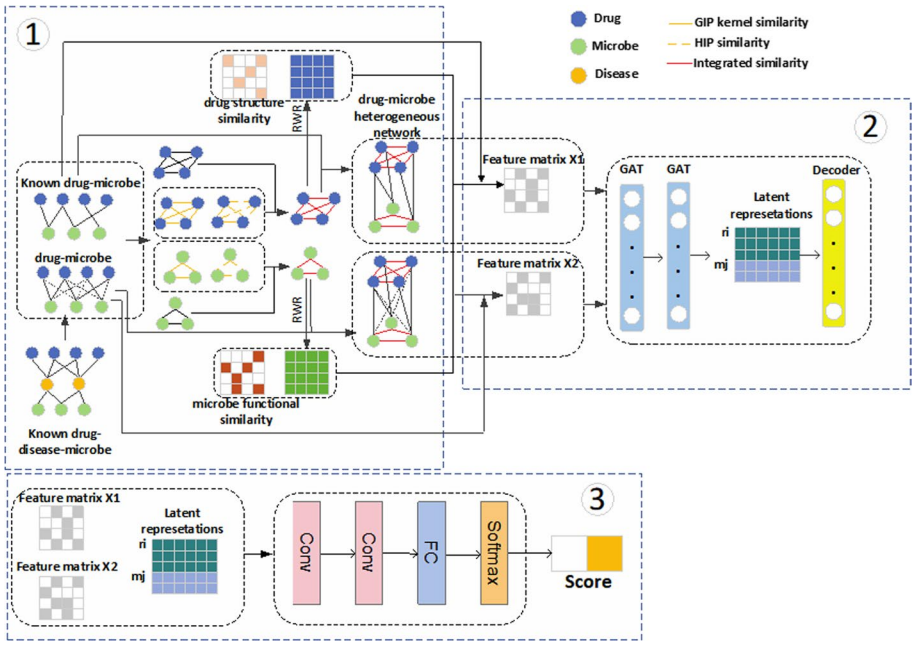

Fig. 1 Flowchart of the GACNNMDA

Part 1: In this part, through adopting multiple measures of similarity, two heterogenous networks HN_1 and HN_2 will be constructed based on downloaded known microbe-drug associations, drug-drug interactions and microbe-microbe interactions.

Part 2: In this part, two feature matrices will be obtained for microbes and drugs by leveraging various attributes of microbes and drugs first, and then, through taking these two feature matrices and two heterogeneous networks as inputs, a two-layer graph attention network will be further designed to learn low dimensional feature representations for microbes and drugs.

Part 3: In this part, a CNN-based classifier will be introduced to calculate possible scores of drug-microbe associations, in which, those newly learned low dimensional feature representations will be integrated with those two feature matrices to form its inputs.

Construction of two heterogeneous networks

For convenience, let n_r and n_m represent the numbers of those newly downloaded drugs and microbes separately. Firstly, based on those newly downloaded known microbe-drug associations, we can obtain a microbe-drug adjacency matrix $A^1 \in R^{n_r * n_m}$ as follows: for any given drug r_i and microbe m_j, if there is a known association between them, then there is $A^1(i,j) = 1$, otherwise there is $A^1(i,j) = 0$.

Secondly, based on those newly downloaded known microbe-drug, microbe-disease and drug-disease associations, we can obtain another microbe-drug adjacency matrix $A^2 \in R^{n_r * n_m}$ as follows: for any given drug r_i, microbe m_j and disease d_k, if there is a known association between r_i and d_k, and a known association between m_j and d_k, simultaneously, then there is $A^2(i,j) = 1$, otherwise there is $A^2(i,j) = A^1(i,j)$.

Finally, based on above matrices A^1 and A^2, we can construct two heterogeneous networks HN_1 and HN_2 respectively according to the methods proposed in the following "Calculation of the Gaussian interaction profile (GIP) kernel similarity for microbes

and drugs" to "Calculation of the Gaussian interaction profile (GIP) kernel similarity for microbes and drugs" sections.

Let $A^v(r_i)$ and $A^v(m_j)$ denote the i-th row and the j-th column of A^v ($v=1,2$) respectively, and $\|\bullet\|$ represent the Frobenius norm, then for any two given drugs r_i and r_j, we can calculate the GIP kernel similarity between them as follows:

$$S_{rg}^v(r_i, r_j) = exp\left(-\gamma 1\left|\left|A^v(r_i) - A^v(r_j)\right|\right|^2\right) \tag{1}$$

$$\gamma 1 = 1/\left(\frac{1}{n_r}\sum_{i=1}^{n_r}\left|\left|A^v(r_i)\right|\right|^2\right) \tag{2}$$

According to above equations, it is easy to see that we can obtain a new GIP kernel similarity matrix $S_{rg}^v \in R^{n_r * n_r}$.

Similarly, for any two given microbes m_i and m_j, we can calculate the GIP kernel similarity between them as follows:

$$S_{mg}^v(m_i, m_j) = exp\left(-\gamma 2\left|\left|A^v(m_i) - A^v(m_j)\right|\right|^2\right) \tag{3}$$

$$\gamma 2 = 1/\left(\frac{1}{n_m}\sum_{i=1}^{n_m}\left|\left|A^v(m_i)\right|\right|^2\right) \tag{4}$$

According to above equations, it is obvious that we can obtain a new GIP kernel similarity matrix $S_{mg}^v \in R^{n_m * n_m}$.

Calculation of the Hamming interaction profile (HIP) similarity for microbes and drugs

Based on the assumption that two nodes will have lower similarity when their interaction profiles are more different. Let $|\bullet|$ denote the number of elements in the profile, then for any two given drugs r_i and r_j, we can calculate the HIP similarity between them as follows:

$$S_{rh}^v(r_i, r_j) = 1 - \frac{\left|A^v(r_i)! = A^v(r_j)\right|}{|A^v(r_i)|} \tag{5}$$

where $\left|A(r_i)! = A(r_j)\right|$ denotes the number of different elements between the profiles $A(r_i)$ and $A(r_j)$.

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

$$S_{mh}^v(m_i, m_j) = 1 - \frac{\left|A^v(m_i)! = A^v(m_j)\right|}{|A^v(m_i)|} \tag{6}$$

where $\left|A(m_i)! = A(m_j)\right|$ denotes the number of different elements between the profiles $A(m_i)$ and $A(m_j)$.

According to above equations, it is obvious that we can obtain two new HIP similarity matrices $S_{rh}^v \in R^{n_r * n_r}$ and $S_{mh}^v \in R^{n_m * n_m}$ separately.

Integrated similarity

Based on S_{rg}^{v}, S_{rh}^{v} and newly downloaded known drug-drug interactions, for any two given drugs r_i and r_j, we can calculate an integrated similarity between them as follows:

$$S_r^v(r_i, r_j) = \begin{cases} 1: & if\ there\ is\ a\ known\ association\ between\ r_i\ and\ r_j \\ \frac{S_{rg}^v(r_i,r_j)+S_{rh}^v(r_i,r_j)}{2}: & otherwise \end{cases} \tag{7}$$

In the same way, based on S_{mg}^{v}, S_{mh}^{v} and newly downloaded known microbe-microbe interactions, for any two given microbes m_i and m_j, we can calculate an integrated similarity between them as follows:

$$S_m^v(m_i, m_j) = \begin{cases} 1: & if\ there\ is\ a\ known\ association\ between\ m_i\ and\ m_j \\ \frac{S_{mg}^v(m_i,m_j)+S_{mh}^v(m_i,m_j)}{2}: & otherwise \end{cases} \tag{8}$$

Hence, based on above newly obtained matrices, we can obtain two new matrices $H^1 \in R^{(n_r+n_m)*(n_r+n_m)}$ and $H^2 \in R^{(n_r+n_m)*(n_r+n_m)}$ as follows:

$$H^1 = \begin{bmatrix} S_r^1 & A^1 \\ (A^1)^T & S_m^1 \end{bmatrix} \tag{9}$$

$$H^2 = \begin{bmatrix} S_r^2 & A^2 \\ (A^2)^T & S_m^2 \end{bmatrix} \tag{10}$$

Obviously, according to above two matrices H^1 and H^2, we can easily construct two heterogeneous networks HN_1 and HN_2 respectively.

Low dimensional feature representations learning for microbes and drugs based on the graph attention network

Construction of two feature matrices

In this section, for any two given drugs r_i and r_j, we would first adopt SIMCOMP2 [26] to calculate the structural similarity between them, as a result, we can obtain a new drug structural similarity matrix S_{rc}. And at the same time, for any two given microbes m_i and m_j, we would adopt the method proposed by Kamneva et al. [27] to calculate the functional similarity between them, as a result, we can obtain a new microbe functional similarity matrix S_{mf} as well.

Moreover, we would further implement a random walk with restart (RWR) on S_r^v and S_m^v to obtain the topological attributes S_{rr}^v, S_{mm}^v of drugs and microbes separately, where the RWR was defined as follows:

$$p_i^{l+1} = 0.1 * Mp_i^l + 0.9 * \varepsilon_i \tag{11}$$

$$\varepsilon_{ij} = \begin{cases} 1 if i = j \\ 0 otherwise \end{cases} \tag{12}$$

Here, p_i^l denotes the probabilities that node i reaches other nodes at the time slot l. M is the transition probability matrix and $\varepsilon_i \varepsilon R^{1*n}$ represents the initial probability vector of node i.

Different from the usual weighted addition of various attribute vectors of nodes to form the feature matrix, we spliced various attributes together to retain more original features. The feature matrices $X^v \in R^{(n_r+n_m)*k_1}$ for two heterogeneous networks were defined as follows:

$$F_r^v = \left[S_{rc}; A^v; S_{rr}^v; A^v\right] \tag{13}$$

$$F_m^v = \left[\left(A^v\right)^T; S_{mf}; \left(A^v\right)^T; S_{mm}^v\right] \tag{14}$$

$$X^v = \begin{bmatrix} F_r^v \\ F_m^v \end{bmatrix} \tag{15}$$

where k_1 denotes the dimension of the feature matrices X^v.

The structure of the graph attention network

Encoder: Firstly, for any given node i in $H^v (v = 1, 2)$, the coefficient of similarity between it and its neighbors would be calculated as follows:

$$e_{ij} = LeakyRelu\left(a\left[W^v X^v(i); W^v X^v\left(j\right)\right]\right), j \in \Phi_i^v \tag{16}$$

$$LeakyRelu(x) = \begin{cases} x & x > 0 \\ \mu x & otherwise \end{cases} \tag{17}$$

Here, $X^v(i)$ denotes the ith row of X^v and a represents a feature mapping operation. W^v is a trainable weight matrix parameter and Φ_i^v is the set of neighbor nodes of node i in H^v, μ is the hypermeter.

Subsequently, the attention score λ_{ij} between node i and node j would be calculated based on e_{ij} according to the following formula:

$$\lambda_{ij} = \frac{exp\left(e_{ij}\right)}{\sum_{k \in \Phi v_i} exp(e_{ik})} \tag{18}$$

Finally, the features would be weighted and summed according to the calculated attention score to obtain the new feature representation of node i as follows:

$$X^v(i)^{'} = Relu\left(\sum_{j \in \Phi v_i} \lambda_{ij} W^v X^v\left(j\right)\right) \tag{19}$$

$$Relu(x) = \begin{cases} x & x > 0 \\ 0 & otherwise \end{cases} \tag{20}$$

After obtaining new feature representations of all nodes in H^v, it is easy to see that we can construct a feature representation matrix $Y^v = \begin{bmatrix} R_r^v \\ R_m^v \end{bmatrix} \in R^{(n_r+n_m)*k_2}$.

Where k_2 denotes the dimension of the feature representation matrix Y^v.

Decoder: The decoder runs an inner product based on newly learned feature representation matrix Y^v as follows:

$$Y^{v\prime} = Sigmoid\left(Y^v \cdot \left(Y^v\right)^T\right) \tag{21}$$

$$Sigmoid(x) = \frac{1}{1+e^{-x}} \tag{22}$$

Optimization

Considering the reconstructed matrix should be as similar as possible to the original matrix, we adopted the MSE loss function to compute the mean of the sum of squares of the differences between $Y^{v\prime}$ and H^v as follows:

$$Loss = \frac{1}{n_r+n_m}\sum_{i=1}^{n_r+n_m}\left|\left|Y^{v\prime}(i) - H^v(i)\right|\right|^2 \tag{23}$$

where $Y^{v\prime}(i)$ and $H^v(i)$ denote the i-th row of $Y^{v\prime}$ and H^v respectively. During training, we used the Adam optimizer to optimize the loss function.

Construction of the CNN-based classifier

In this section, we treated the microbe-drug association prediction as a binary classification problem and designed a classifier based on the convolutional neural network to calculate possible scores of potential drug-microbe associations. For the input of the classifier, we first constructed two new feature matrices N_r^v and N_m^v for drugs and microbes separately as follows:

$$N_r^v = \left[R_r^v; F_r^v\right] \tag{24}$$

$$N_m^v = [R_m^v; F_m^v] \tag{25}$$

And then, let k_3 denote the dimension of the new feature matrix, then for any given drug r_i and microbe m_j, the feature matrix $F^v(i,j) = \begin{bmatrix} N_r^v(i) \\ N_m^v(j) \end{bmatrix} \in R^{2*k_3}$ would be fed into the classifier to calculate the score between i and j. Here,$N_r^v(i)$ and $N_m^v(j)$ denote the i-th and the j-th row of N_r^v and N_m^v, respectively.

In the convolutional layer, we adopted zero padding to enlarge the edges and set the size of the convolution kernel to 3×3. The convolutional operation in the i-th layer were defined as follows:

$$F_i = Relu(F_{i-1} \otimes G_i + b_i) \tag{26}$$

where $\otimes$ represents the operation of convolution, G_i is the weight matrix, and b_i is the offset vector. It is worth mentioning that we added the BatchNorm2d [28] to normalize data to enhance performance stability before *Relu*.

After inputs having gone through two convolution layers, it would be flattened into a vector. And then, a full-connected layer and a softmax layer would be used to obtain scores of two associative categories, based on which, we would adopt scores of the second category as predicted scores of potential microbe-drug associations in GACNNMDA. Obviously, based on H^1 and H^2, we can obtain two score matrices $Score^1$ and $Score^2$ respectively. Hence, a final score matrix $Score \in R^{n_r * n_m}$ can be calculated as follows:

$$Score(i,j) = \frac{Score^1(i,j) + Score^2(i,j)}{2} \tag{27}$$

Moreover, in the classifier, we utilized the cross-entropy as loss function and Adam optimizer to minimize the loss function. Here, the loss function L^v (v=1, 2) was defined as follows:

$$L^v = -\frac{1}{n_r * n_m} \sum a_{ij}^v logs_{ij}^v + \left(1 - a_{ij}^v\right) log\left(1 - s_{ij}^v\right) \tag{28}$$

where a_{ij}^v and s_{ij}^v represent the ij-th entry of A^v and $Score^v$ respectively.

Results

Comparison with state-of-the-art methods

Considering that there are few computational methods and codes available for microbial-drug association prediction, we compared GACNNMDA with four existing microbe-drug association prediction methods such as HMDAKATZ [19], GCNMDA [22], EGATMDA [23] and Graph2MDA [24], and two methods for link prediction problems in the bioinformatics field such as LAGCN [29] and NTSHMDA [30]. Among them, LAGCN [29] is a graph convolutional network with attention mechanism based method designed to infer unknown drug-disease associations. NTSHMDA [30] is a model based on random walk with restart for predicting microbe-disease associations.

During experiments, we settled with original parameters for all these competitive methods and ran them on the well-known public database MDAD for a fair comparison. In addition, we adopted the framework of fivefold cross validation (CV) to evaluate these methods, in which, 20% of known associations and 20% of unknown associations would be randomly selected as the testing set, and the remaining 80% of known associations and unknown associations as the training set [31]. And then, we selected the AUC, AUPR, Accuracy and F1-Score as the metrics of performance evaluation. Experimental results were shown in Table 2. Due to the incomplete code proposed by Deng et al. [24], we directly referenced the results in Graph2MDA. As a result, the ROC and PR curves were drawn in Figs. 2 and 3 separately, in which, those evaluation metrics are calculated as follows:

$$TPR = \frac{TP}{TP + FN} \tag{29}$$

Table 2 The AUCs, AUPRs, accuracy and F1-scores achieved by compared methods based on MDAD under fivefold CV

Methods	AUC	AUPR	Accuracy	F1-score
HMDAKATZ	0.8712 ± 0.0010	0.2327 ± 0.0068	0.9774	0.3546
GCNMDA	0.9427 ± 0.0002	0.9133 ± 0.0031	0.9905	0.6672
EGATMDA	0.9585 ± 0.0053	0.9268 ± 0.0142	0.9081	0.6871
Graph2MDA	0.9567 ± 0.0039	**0.9380 ± 0.0098**	0.9934	0.7091
LAGCN	0.8533 ± 0.0070	0.3571 ± 0.0051	0.9413	0.0423
NTSHMDA	0.8483 ± 0.0020	0.1892 ± 0.0056	0.9896	0.1838
GACNNMDA	**0.9777 ± 0.0109**	0.7015 ± 0.0366	**0.9945**	**0.7091**

Bold values indicate the best results achieved by all these competitive methods

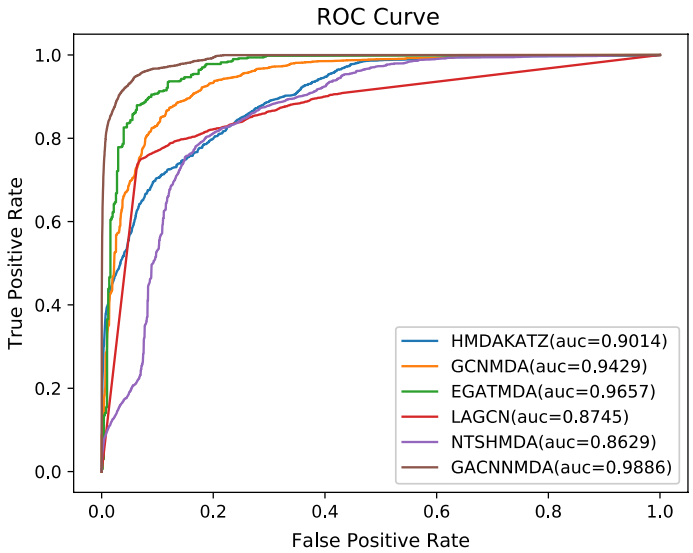

Fig. 2 The ROC curves of six competitive methods

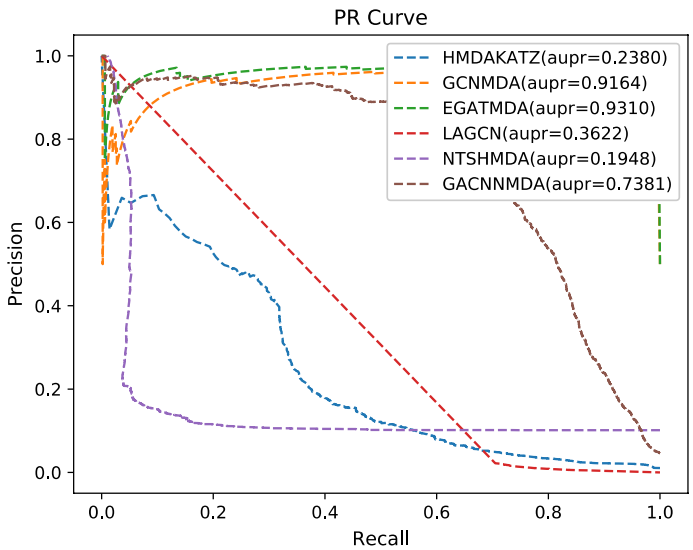

Fig. 3 The PR curves of six competitive methods

$$FPR = \frac{FP}{TN + FP} \tag{30}$$

$$Precision = \frac{TP}{TN + FP} \tag{31}$$

$$Recall = \frac{TP}{TP + FN} \tag{32}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{33}$$

$$F1 - score = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{34}$$

Here, TP and TN represent the numbers of positive and negative samples predicted correctly, respectively. FN and FP denote the numbers of positive and negative samples that are incorrectly identified, separately.

As shown in Table 2, it is obvious that GACNNMDA can achieve the highest AUC value of 0.9777 ± 0.0109, which is 2.57% higher than the second highest AUC value of 0.9585 ± 0.0053 obtained by EGATMDA. For evaluation metrics of accuracy and f1-score, GACNNMDA can also achieve the highest values of 0.9945 and 0.7091 respectively. Although in terms of AUPR value, GACNNMDA can only outperform half of all these competitive methods, we can say that GACNNMDA is an effective tool for potential microbe-drug association prediction.

Hyperparameter sensitivity analysis

Considering that there are several hyperparameters in GACNNMDA, including the learning rate of GAT, the dropout of GAT and the learning rate of CNN, therefore, in this section, we would perform a fivefold CV on the MDAD dataset for 10 times and observe the average AUC value to tune the values of these parameters.

For convenience, let $lr1$, dp and $lr2$ denote the learning rate of GAT, the dropout of GAT and the learning rate of CNN respectively. During the tuning process, we first tested the values of $lr1$ in the range of {0.0001, 0.001, 0.01, 0.05, 0.1} and illustrated experimental results in Fig. 4a. As shown in Fig. 4a, GACNNMDA achieved the best performance when $lr1$ was set to 0.001. And then, we limited the values of dp in the range of {0.2, 0.4, 0.5, 0.7} and illustrated experimental results in Fig. 4b. From observing Fig. 4b, it is easy to see that the most suitable value of dp is 0.4. Finally, we restricted the values of $lr2$ in {0.0001, 0.001, 0.01, 0.05, 0.1} and showed experimental results in Fig. 4c. As illustrated in Fig. 4c, when $lr2$ was set to 0.001, the performance of GACNNMDA would be the best.

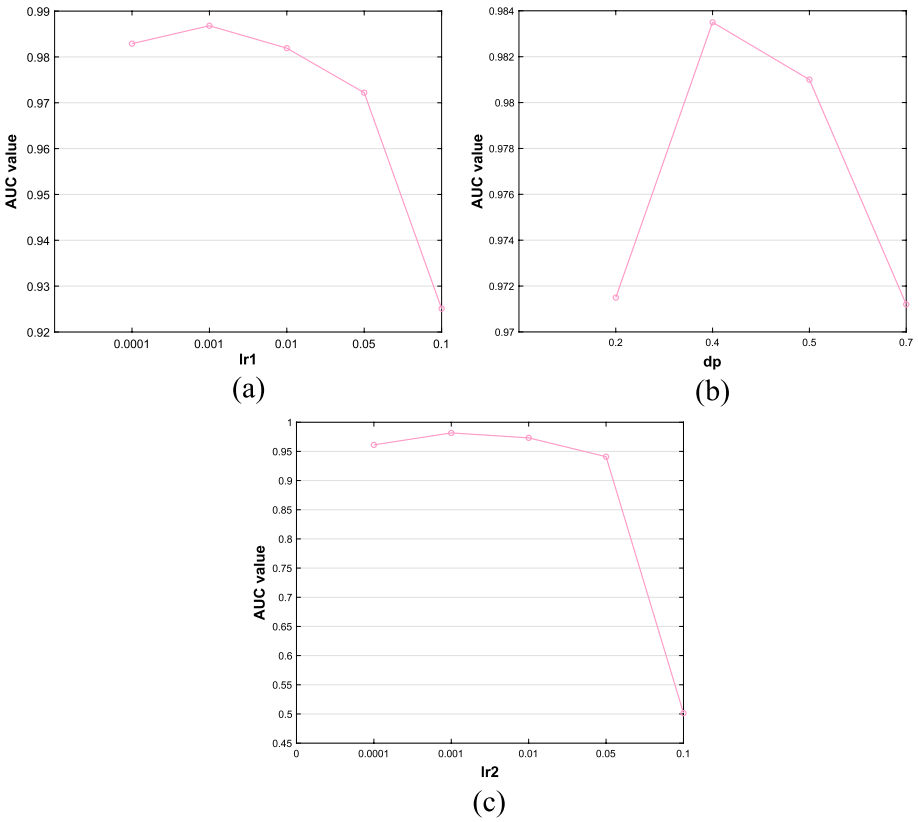

Fig. 4 Analysis of the impact of hyperparameters on performance of GACNNMDA. The subfigures from (**a**) to (**c**) show the AUC values of related values of the learning rate of GAT, the dropout of GAT and the learning rate of CNN, respectively

Case studies

In order to further demonstrate the prediction performance of GACNNMDA, case studies on two popular drugs and two microbes will be done in this section. And in experiments of case studies, the top 20 microbes or drugs inferred by GACNNMDA based on the database of MDAD will be selected out for investigation first, and then, we will search published PubMed literatures to verify whether these predicted candidates having been reported by existing references.

The first drug that we chose for case studies is Ciprofloxacin, which is a fluorinated quinolone antibiotic, and a large number of studies have shown that it is associated with a wide range of human microbes [32]. For instance, Paul et al. found that Amphotericin-B and 5% ciprofloxacin can effectively hindered the growth of Pseudomonas aeruginosa and Candida albicans [33]. *Staphylococcus aureus*, *Staphylococcus epidermidis*, *Bacillius subtilius*, *Escherichia coli* and Mycobacterium tuberculosis are susceptible to Ciprofloxacin [34]. The second drug that we chose for case studies is Moxifloxacin, which is a fluoroquinolone antibiotic [35], and has been proven to be associated with antibiotic-resistant bacteria (ARB) [36] and Listeria monocytogenes [37]. And as a result, we illustrated the top 20 predicted ciprofloxacin-associated and moxifloxacin-associated microbes in Tables 3 and 4 respectively. From

Table 3 The top 20 candidate Ciprofloxacin-associated microbes

Microbe	Evidence	Microbe	Evidence
Staphylococcus aureus	PMID: 32488138	*Streptococcus sanguinis*	PMID: 21507381
Mycobacterium tuberculosis	PMID: 30020039	*Enterococcus faecalis*	PMID: 27790716
Escherichia coli	PMID: 26607324	*Eggerthella lenta*	Unconfirmed
Bacillus subtilis	PMID: 33218776	*Salmonella enterica*	PMID: 6933017
Haemophilus influenzae	PMID: 27292570	Human herpesvirus 5	Unconfirmed
Stenotrophomonas maltophilia	PMID: 14982788	*Propionibacterium acnes*	PMID: 25445201
Pseudomonas aeruginosa	PMID: 33875431	*Klebsiella pneumoniae*	PMID: 27257956
Morganella morganii	PMID: 29942700	*Staphylococcus cohnii*	PMID: 19780489
Providencia stuartii	PMID: 1337751	*Serratia marcescens*	PMID: 2071875
Proteus vulgaris	PMID: 34638966	*Staphylococcus epidermis*	PMID: 10632381

The top 10 predicted microbes are included in the first column, while the top 11–20 predicted microbes are included in the third column records

Table 4 The top 20 candidate Moxifloxacin-associated microbes

Microbe	Evidence	Microbe	Evidence
Bacillus subtilis	PMID: 30036828	*Staphylococcus Aureus*	PMID: 31689174
Haemophilus influenzae	PMID: 11856249	Enterococcus faecium	PMID: 10629010
Stenotrophomonas maltophilia	PMID: 27257956	Human herpesvirus 5	Unconfirmed
Candida albicans	PMID: 21108571	*Proteus vulgaris*	Unconfirmed
Mycobacterium avium	PMID: 21353489	Bacillus cereus	PMID: 21834669
Pseudomonas aeruginosa	PMID: 31691651	*Streptococcus pneumoniae*	PMID: 31542319
Campylobacter jejuni	PMID: 16027651	Serratia marcescens	PMID: 34439014
Staphylococcus aureus	PMID: 31689174	*Streptococcus mutans*	PMID: 29160117
Neisseria gonorrhoeae	PMID: 26603424	*Klebsiella pneumoniae*	PMID: 33406110
Escherichia coli	PMID: 31542319	Bacteroides	PMID: 18385145

The top 10 predicted microbes are included in the first column, while the top 11–20 predicted microbes are included in the third column records

observing Tables 3 and 4, it is easy to see that there are 18 and 17 out of top 20 predicted microbes having been validated by existing literatures separately.

Besides, the first microbe that we chose for case studies is HIV-1 (Human Immunodeficiency Virus type 1), which is the cause of the acquired immunodeficiency syndrome (AIDS). There are many drugs associated with HIV-1. For example, Viani et al. found that long-term zalcitabine for treating HIV-1 phenotypes in children is useful [38]. Chong et al. proved that combination of delavirdine, zidovudine and didanosine can inhibit the growth of the HIV-1 [39]. The second microbe that we chose for case studies is mycobacterium tuberculosis, which is the cause of the pulmonary tuberculosis [40]. And as a result, we showed the top 20 predicted HIV-1-associated and mycobacterium tuberculosis-associated drugs in Tables 5 and 6 respectively. From observing Tables 5 and 6, it is obvious that there are 18 and 15 out of top 20 predicted drugs having been verified by existing literatures. Hence, we can draw a conclusion that GACNNMDA can achieve satisfactory prediction performance in both case studies of microbes and drugs.

Table 5 The top 20 candidate Human immunodeficiency virus type 1-associated drugs

Drug	Evidence	Drug	Evidence
Zalcitabine	PMID: 9498433	Cala + B20nolide A	PMID: 8930168
Abacavir	PMID: 11797183	Tenofovir	PMID: 33336698
Fosamprenavir	PMID: 19515730	Bevirimat	PMID: 19024627
Didanosine	PMID: 9107385	Dolutegravir	PMID: 31865558
Indinavir	PMID: 8970946	Peptide 1037	Unconfirmed
Delavirdine	PMID: 9107385	Vancomycin	Unconfirmed
Tipranavir	PMID: 17360759	Nevirapine	PMID: 20384494
Stavudine	PMID: 8568296	Enfuvirtide	PMID: 14523775
Atazanavir	PMID: 15585441	Lopinavir	PMID: 20836579
Zidovudine	PMID: 2012453	Trimethoprim-sulfamethoxazole	PMID: 9142796

The top 10 predicted drugs are included in the first column, while the top 11–20 predicted drugs are included in the third column records

Table 6 The top 20 candidate Mycobacterium tuberculosis-associated drugs

Drug	Evidence	Drug	Evidence
Ciprofloxacin	PMID: 16270765	Meropenem	PMID: 22906310
Aminosalicylic acid	PMID: 26033719	Polysorbate 80	Unconfirmed
SQ109	PMID: 22258923	Pyrogallol	PMID: 13411428
Colistin	PMID: 26183185	Pefloxacin	PMID: 1909062
Ethambutol	PMID: 27806932	Zinc oxide	PMID: 33845951
Tobramycin	PMID: 19723387	Desipramine	PMID: 7649718
Pyrazinamide	PMID: 26521205	Saquinavir	PMID: 33841429
Telithromycin	unconfirmed	Gatifloxacin	PMID: 17267339
Capreomycin	PMID: 29311078	Undecanoic acid	Unconfirmed
Trans-2-nonenal	Unconfirmed	Piperacillin-Tazobactam	Unconfirmed

The top 10 predicted drugs are included in the first column, while the top 11–20 predicted drugs are included in the third column records

Conclusion and discussion

In this paper, we presented a novel calculation method named GACNNMDA, an integrated framework of GAT-based autoencoder and CNN-based classifier, for prediction of potential microbe-drug associations. The main contributions of our model include the following three points.

1. We introduced known microbe-disease-drug associations into the predictive model and made up for the sparsity of known microbe-drug associations to some extent.
2. For the inputs of GAT and CNN, we spliced multiple attributes of microbes and drugs together to form two feature matrices, which can retain more original features of microbes and drugs. Hence, more useful information can be learned by the GAT and the CNN.
3. Compared with existing state-of-the-art methods for predicting potential microbe-drug associations, our model can achieve better performance.

However, there is still room to improve our prediction model. In the feature, we can leverage more biological information, such as microbe sequences [24] and

side-effect-based drug similarity [41]. Additionally, for those attributes of microbes and drugs used in GACNNMDA, we can make an assessment of their importance to better use each kind of attribute and further improve the performance of our model. Finally, we can design a new activation to improve the training speed of GAT and CNN such as Li et al. [42].

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12859-023-05158-7.

Additional file 1: The newly downloaded dataset of diseases.

Additional file 2: The newly downloaded dataset of drugs.

Additional file 3: The newly downloaded dataset of microbes.

Additional file 4: The newly downloaded dataset of drug-disease associations.

Additional file 5: The newly downloaded dataset of drug-drug interactions.

Additional file 6: The newly downloaded dataset of microbe-drug associations.

Additional file 7: The newly downloaded dataset of microbe-disease associations.

Additional file 8: The newly downloaded dataset of microbe-microbe interactions.

Acknowledgements
Not applicable

Author contributions
LW supervised the study. QM and YQT designed the model and conducted the experiments, YQT and QM wrote this paper. LW provide suggestions and revised the manuscript. All authors read and approved the final manuscript.

Funding
This work was partly sponsored by the National Natural Science Foundation of China (No. 62272064) and the Key project of Changsha Science and technology Plan (No. KQ2203001).

Data availability
The data and code can be found online at: https://github.com/tyqGitHub/TYQ/tree/master/ GACNNMDA.

Declarations

Ethical approval and consent to participate
Not applicable.

Consent to publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Received: 18 November 2022 Accepted: 24 January 2023
Published online: 02 February 2023

References

1. Huttenhower C, Gevers D, Knight R, et al. Structure, function and diversity of the healthy human microbiome. Nature. 2012;486(7402):207–14.
2. Ventura M, O'Flaherty S, Claesson MJ, et al. Genome-scale analyses of health-promoting bacteria: probiogenomics. Nat Rev Microbiol. 2009;7(1):61–71.
3. Sommer F, Backhed F. The gut microbiota masters of host development and physiology. Nat Rev Microbiol. 2013;11(4):227–38.
4. Kau AL, Ahern PP, Griffin NW, et al. Human nutrition, the gut microbiome and the immune system. Nature. 2011;474(7351):327–36.
5. Sprockett D, Fukami T, et al. Role of priority effects in the early-life assembly of the gut microbiota. Nat Rev Gastroenterol Hepatol. 2018;15(4):197–205.
6. Ximenez C, Torres J. Development of microbiota in infants and its role in maturation of gut mucosa and immune system. Arch Med Res. 2017;48(8):666–80.
7. Tilg HA, et al. The intestinal microbiota in colorectal cancer. Cancer Cell. 2018;33(6):954–64.
8. Cani PD, et al. Novel insight into the role of microbiota in colorectal surgery. Gut J Br Soc Gastroenterol. 2017;66(4):738–49.

9. Routy B, Gopalakrishnan V, et al. The gut microbiota influences anticancer immunosurveillance and general health. Nat Rev Clin Oncol. 2018;15(6):382–96.
10. Shanahan F, Sinderen DV, O'Toole PW, et al. Feeding the microbiota: transducer of nutrient signals for the host. Gut. 2017;66(9):1709–17.
11. Cremonesi E, Governa V, Garzon JFG, Mele V, Amicarella F. Gut microbiota modulate T cell trafficking into human colorectal cancer. Gut J Br Soc Gastroenterol. 2018;67(11):1984–94.
12. Ogino S, Nowak JA, Hamada T, et al. Integrative analysis of exogenous, endogenous, tumour and immune factors for precision medicine. Gut. 2018;67(6):1168–80.
13. Xiang Y-T, Li W, Zhang Q, et al. Timely research papers about COVID-19 in China. Lancet. 2020;395(10225):684–5.
14. McCoubrey LE, Gaisford S, Orlu M, et al. Predicting drug-microbiome interactions with machine learning. Biotechnol Adv. 2022;54: 107797.
15. Zimmermann M, Zimmermann-Kogadeeva M, Wegmann R, et al. Mapping human microbiome drug metabolism by gut bacteria and their genes. Nature. 2019;570(7762):462–7.
16. Sun YZ, Zhang DH, et al. MDAD: a special resource for microbe-drug associations. Front Cell Infect Microbiol. 2018.
17. Rajput A, Thakur A, Sharma S, et al. aBiofilm: a resource of anti-biofilm agents and their potential implications in targeting antibiotic drug resistance. Nucleic Acids Res. 2018;46(D1):D894–900.
18. Pia A, Ai A, Hl A, et al. Discovery and development of safe-in-man broad-spectrum antiviral agents. Int J Infect Dis. 2020;93:268–76.
19. Zhu L, Duan G, Yan C, et al. Prediction of microbe-drug associations based on KATZ measure. In 2019 IEEE international conference on bioinformatics and biomedicine (BIBM) 2019. pp. 183–187.
20. Long Y, Luo J. Association mining to identify microbe drug interactions based on heterogeneous network embedding representation. IEEE J Biomed Health Inform. 2021;25(1):266–75.
21. Zhu L, Wang J, Li G, et al. Predicting microbe-drug association based on similarity and semi-supervised learning. Am J Biochem Biotechnol. 2021;17(1):50–8.
22. Long Y, Wu M, Keong KC, et al. Predicting human microbe–drug associations via graph convolutional network with conditional random field. Bioinformatics. 2020;36(19):4918–27.
23. Long Y, Wu M, Liu Y, et al. Ensembling graph attention networks for human microbe–drug association prediction. Bioinformatics. 2020;36(Supplement 2):i779–86.
24. Deng L, Huang Y, Liu X, et al. Graph2MDA: a multi-modal variational graph embedding model for predicting microbe–drug associations. Bioinformatics. 2022;38(4):1118–25.
25. Wang L, Tan Y, Yang X, et al. Review on predicting pairwise relationships between human microbes, drugs and diseases: from biological data to computational models. Brief Bioinf. 2022;23(3):bbac080.
26. Hattori M, Tanaka N, Kanehisa M, et al. SIMCOMP/SUBCOMP: chemical structure search servers for network analyses. Nucleic Acids Res. 2010;38(2):W652–6.
27. Kamneva OK. Genome composition and phylogeny of microbes predict their co-occurrence in the environment. PLoS Comput Biol. 2017;13(2): e1005366.
28. Ioffe S, Szegedy C. Batch normalization: accelerating deep network training by reducing internal covariate shift. In Proceedings of the 32nd international conference on machine learning, vol 37; 2015. pp. 448–456.
29. Yu Z, Huang F, Zhao X, et al. Predicting drug–disease associations through layer attention graph convolutional network. Brief Bioinf. 2020;22(4):bbaa243.
30. 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 Bioinf. 2020;17(4):1341–51.
31. Cai L, Lu C, Xu J, et al. Drug repositioning based on the heterogeneous information fusion graph convolutional network. Brief Bioinf. 2021;22(6):bbab319.
32. Campoli-Richards DM, Monk JP, Price A, et al. Ciprofloxacin. Drugs. 1988;35(4):373–447.
33. Paul D, Saha S, Singh N, et al. Successful control of a co-infection caused by *Candida albicans* and *Pseudomonas aeruginosa* in Keratitis. Infect Disord Drug Targets. 2021;21(2):284–8.
34. Castro W, Navarro M, Biot C. Medicinal potential of ciprofloxacin and its derivatives. Future Med Chem. 2013;5(1):81–96.
35. Balfour JAB, et al. Moxifloxacin. Drugs. 1999;57(3):363–73.
36. Loyola-Rodriguez JP, Ponce-Diaz ME, Loyola-Leyva A, et al. Determination and identification of antibiotic-resistant oral streptococci isolated from active dental infections in adults. Acta Odontol Scand. 2018;76(4):229–35.
37. Tahoun ABMB, Abou Elez RMM, Abdelfatah EN, et al. Listeria monocytogenes in raw milk, milking equipment and dairy workers: Molecular characterization and antimicrobial resistance patterns. J Glob Antimicrobial Res. 2017;10:264–70.
38. Viani RM, Smith IL, Spector SA. Human immunodeficiency virus type 1 phenotypes in children with advanced disease treated with long-term zalcitabine. J Infect Dis. 1998;177(3):565–70.
39. Chong K-T, Pagano PJ. Inhibition of human immunodeficiency virus type 1 infection in vitro by combination of delavirdine, zidovudine and didanosine. Antiviral Res. 1997;34(1):51–63.
40. Koch A, Mizrahi V. Mycobacterium tuberculosis. Trends Microbiol. 2018;26(6):555–6.
41. Kuhn M, Campillos M, Letunic I, et al. A side effect resource to capture phenotypic effects of drugs. Mol Syst Biol. 2010;6(1):343.
42. Li H, Wang Y, Zhang Z, et al. Identifying microbe-disease association based on a novel back-propagation neural network model. IEEE ACM Trans Comput Biol Bioinf. 2021;18(6):2502–13.

Publisher's Note