BMC Bioinformatics

RESEARCH

Open Access

GSAMDA: a computational model for predicting potential microbe–drug associations based on graph attention network and sparse autoencoder

Yaqin Tan[1,2], Juan Zou[1], Linai Kuang[1], Xiangyi Wang[3], Bin Zeng[2,3], Zhen Zhang[3] and Lei Wang[1,2,3*]

*Correspondence:
wanglei@xtu.edu.cn

[1] Key Laboratory of Hunan Province for Internet of Things and Information Security, Xiangtan University, Xiangtan 411105, China
[2] Institute of Bioinformatics Complex Network Big Data, Changsha University, Changsha 410022, China
[3] Big Data Innovation and Entrepreneurship Education Center of Hunan Province, Changsha University, Changsha 410022, China

Abstract

Background: Clinical studies show that microorganisms are closely related to human health, and the discovery of potential associations between microbes and drugs will facilitate drug research and development. However, at present, few computational methods for predicting microbe–drug associations have been proposed.

Results: In this work, we proposed a novel computational model named GSAMDA based on the graph attention network and sparse autoencoder to infer latent microbe–drug associations. In GSAMDA, we first built a heterogeneous network through integrating known microbe–drug associations, microbe similarities and drug similarities. And then, we adopted a GAT-based autoencoder and a sparse autoencoder module respectively to learn topological representations and attribute representations for nodes in the newly constructed heterogeneous network. Finally, based on these two kinds of node representations, we constructed two kinds of feature matrices for microbes and drugs separately, and then, utilized them to calculate possible association scores for microbe–drug pairs.

Conclusion: A novel computational model is proposed for predicting potential microbe–drug associations based on graph attention network and sparse autoencoder. Compared with other five state-of-the-art competitive methods, the experimental results illustrated that our model can achieve better performance. Moreover, case studies on two categories of representative drugs and microbes further demonstrated the effectiveness of our model as well.

Keywords: Microbe–drug associations, Graph attention network-based autoencoder, Sparse autoencoder

Background

Microorganisms, including bacteria, viruses, archaea, fungi and protozoa, are dynamic, diverse and complex genetic reservoirs that exist in interactive flux, colonize human cells, and play significant roles in human beings [1]. The microbial function is to protect the pathogens, improve and enhance metabolism and immunity capability [2]. For example,

microbes can resist the invasion of opportunistic pathogens [3], promote the synthesis of sugar metabolism and synthesis the necessary vitamins to boost T-cell responses [4]. Maintaining the homeostasis of internal environment of organisms is inseparable from the regulation of microorganisms [5]. Unusual growth or decline of microorganisms will influence human health and cause diseases, such as obesity [6], inflammatory bowel disease [7], and even cancer [8]. For instance, pathogens, including bacteria and viruses, may cause infectious diseases such as the COVID-19 [9]. Also, while using drugs to treat microbe-caused diseases, the microbiome may affect the physiological action of drugs in turn. Several studies have shown that not only microbial metabolism can significantly affect the clinical response to drugs, but also the administration of drugs can similarly affect the microbiome [1, 10, 11]. Hence, uncovering potential associations between microbes and drugs will be helpful for the development of drugs and the treatment of human diseases. Due to the high cost and time-consuming of clinical and biological experiments, it is obvious that effective computational approaches for predicting possible microbe–drug associations will be useful complements of traditional web-lab experiments.

Recently, researchers have published multiple databases such as MDAD [12] and aBiofilm [13], which include a large number of experimentally validated microbe–drug associations. And based on these databases, a series of calculation methods have been proposed to detect latent microbe–drug associations and achieved a certain degree of effects. For instance, Zhu et al. proposed a method called HMDAKATZ by adopting KATZ metric to infer potential microbe–drug associations [14]. Long et al. designed a calculation framework named HNERMDA for possible microbe–drug association prediction through combining metapath2vec with bipartite network recommendation [15]. Furthermore, a computational model called LRLSMDA was proposed in reference [16] for identifying microbe–drug associations based on the Laplacian Regularized Least Square algorithm. Literature [17] introduced a calculation scheme named GCNMDA based on the Graph Convolutional Network (GCN) and Conditional Random Field (CRF) to discover associations between microbes and drugs. In reference [18], a method called EGATMDA was designed based on the framework of graph attention networks to predict possible microbe–drug associations. Additionally, Deng et al. conceived a calculation model named Graph2MDA through applying a variational graph autoencoder to infer microbe–drug associations [19].

Most of the above methods took multiple node features into account and fed them into the same model for prediction. Hence, considering that different node features can be learned by different models may have better performance, we classified node features as topological features and attribute features and learn the representations of these two features through graph attention network(GAT) and sparse autoencoder(SAE) respectively. GAT can propagate the information from local neighbors to learn effective representations and has been widely and successfully used in the field of association prediction such as Long et al. [18], Liu et al. [20]. SAE can extract relatively sparse and useful features by introducing a sparse penalty term on autoencoder [21].

In this paper, we introduced a novel calculation method called GSAMDA based on the graph attention network (GAT) and the sparse autoencoder (SAE) to predict potential microbe–drug associations. In GSAMDA, a heterogeneous network would be

constructed first based on the Gaussian interaction profile (GIP) kernel similarity and Hamming interaction profile (HIP) similarity for microbes and drugs. And then, for each node in the heterogeneous network, a unique topological representation would be learned by adopting a GAT-based autoencoder. Simultaneously, based on multiple features of microbes and drugs, we would further apply SAE to learn a unique attribute representation for each node in the heterogeneous network as well. Thereafter, through combining these two types of node representations with multiple features of microbes and drugs, such as drug structure similarity, microbe functional similarity, drug–disease associations and microbe–disease associations, a unique feature matrix would be built for each node in the heterogeneous network, which would be utilized to obtain predicted scores for possible microbe–drug associations. Finally, in order to verify the prediction performance of GSAMDA, we performed case studies and intensive comparison experiments based on two well-known public databases, and results demonstrated that GSAMDA outperformed five state-of-the-art competitive methods, which means that GSAMDA not only can achieve satisfactory predictive performance, but also may be a kind of useful tool for potential microbe–drug association prediction in the future.

Materials and methods

Data sources

In this manuscript, we first downloaded known microbe–drug associations from two public databases such as MDAD (http://www.chengroup.cumt.edu.cn/MDAD/) and aBiofilm (http://bioinfo.imtech.res.in/manojk/abiofilm/) separately. As a result, we obtained 2470 clinically or experimentally verified microbe–drug associations between 1373 drugs and 173 microbes from the MDAD, while 2884 known microbe–drug associations between 1720 drugs and 140 microbes from the aBiofilm. And then, we further collected known drug–disease associations and known microbe–disease associations from the dataset proposed by Wang et al. [22] as well. During the experiment, only diseases associating with at least one drug and one microbe in the MDAD or aBiofilm, and associations related with these diseases, would be kept. Hence, we finally obtained 109 different diseases, 232 different drugs, 1121 different drug–disease associations and 402 different microbe–disease associations from the MDAD, and 72 different diseases, 103 different drugs, 435 different drug–disease associations and 254 different microbe–disease associations from the aBiofilm. The detailed numbers of these aforementioned data were shown in the following Table 1.

Methods

As shown in Fig. 1, GSAMDA mainly consists of five parts:

Table 1 The detailed numbers of microbes, drugs, diseases and related associations in the MDAD and aBiofilm

Type (MDAD/aBiofilm)	Microbes	Drugs	Diseases	associations
Microbe–drug associations	173/140	1373/1720	–	2470/2884
Microbe–disease associations	73/59	–	109/72	402/254
Drug–disease associations	–	233/103	109/72	1121/435

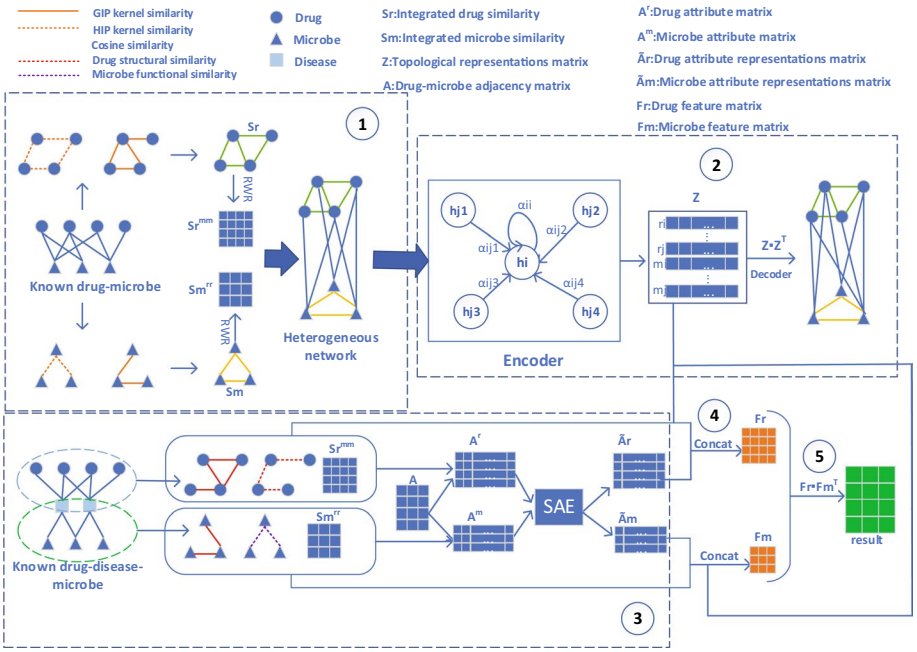

Fig. 1 The overall architecture of GSAMDA

Step 1. Constructing the heterogeneous network *HN* by adopting integrated microbe similarities and drug similarities;

Step 2. Learning topological representations for nodes in *HN* based on the GAT;

Step 3. Learning attribute representations for nodes in *HN* based on the SAE;

Step 4. Constructing feature matrices for nodes in *HN* through combining their topological representations and attribute representations with multiple original attributes of them;

Step 5. Computing possible association scores for microbe–drug pairs based on their feature matrices.

Constructing the heterogeneous network *HN*

In this section, based on newly downloaded drugs, microbes and known microbe–drug associations, we would build the heterogeneous network *HN* as follows.

Firstly, we defined $A \in R^{n_r \times n_m}$ as an adjacency matrix, where n_r and n_m denote the numbers of newly downloaded drugs and microbes separately. In A, for any given drug r_i and microbe m_j, if there is a known association between them, then there is $A_{ij}=1$, otherwise there is $A_{ij}=0$.

Secondly, let $A(r_i)$ and $A(m_j)$ denote the i-th row and the j-th column of A respectively, then for any two given drugs r_i and r_j, we would estimate the GIP kernel similarity $S_r^{GIP}(r_i, r_j) \in R^{n_r \times n_r}$ between them as follows:

$$S_r^{GIP}(r_i, r_j) = exp(-\gamma_r ||A(r_i) - A(r_j)||^2) \tag{1}$$

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

Similarly, for any two given microbes m_i and m_j, we would evaluate the GIP kernel similarity $S_m^{GIP}(m_i, m_j) \in R^{n_m \times n_m}$ between them as follows:

$$S_m^{GIP}(m_i, m_j) = exp(-\gamma_m ||A(m_i) - A(m_j)||^2) \tag{3}$$

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

Here, $||\cdot||$ is the Frobenius norm.

Thirdly, inspired by the work proposed by Xu et al. [23], we further adopted the HIP similarity to measure the similarities between drugs or microbes based on the assumption that two nodes will have lower similarity when their interaction profiles are more different. To be specific, for any two given drugs r_i and r_j, the HIP similarity $S_r^{HIP}(r_i, r_j) \in R^{n_r \times n_r}$ between them would be computed as follows:

$$S_r^{HIP}(r_i, r_j) = 1 - \frac{|A(r_i)! = A(r_j)|}{|A(r_i)|} \tag{5}$$

where $|A(r_i)! = A(r_j)|$ denotes the number of different elements between the profiles $A(r_i)$ and $A(r_j)$, and $|A(r_i)|$ represents the number of elements in $A(r_i)$. Similarly, for any two given microbes m_i and m_j, the HIP similarity $S_m^{HIP}(m_i, m_j) \in R^{n_m \times n_m}$ between them could be estimated as follows:

$$S_m^{HIP}(m_i, m_j) = 1 - \frac{|A(m_i)! = A(m_j)|}{|A(m_i)|} \tag{6}$$

Finally, considering that the values in both the matrices S_r^{GIP} and S_r^{HIP} range from 0 to 1, we could combine these two matrices into a new matrix $S_r \in R^{n_r \times n_r}$ as follows:

$$S_r = (S_r^{GIP} + S_r^{HIP})/2 \tag{7}$$

Similarly, a novel matrix $S_m \in R^{n_m \times n_m}$ could be obtained by integrating S_m^{GIP} and S_m^{HIP} as follows:

$$S_m = (S_m^{GIP} + S_m^{HIP})/2 \tag{8}$$

Thereafter, a matrix $N \in R^{(n_r+n_m)\times(n_r+n_m)}$ could be constructed through combining S_r and S_m with the adjacency matrix A as follows:

$$N = \begin{bmatrix} S_r & A \\ A^T & S_m \end{bmatrix} \tag{9}$$

Here, A^T is the transposed matrix of A.

Obviously, based on above matrix N, we can easily design a heterogeneous network HN consisting of $n_r + n_m$ different nodes, in which, there is an edge between any two nodes i and j, if and only if there is $N(i, j) \neq 0$.

Learning topological representations for nodes in *HN*

The graph attention network (GAT) is an extension of the graph convolution network, it can overcome some shortcomings of graph convolution by using the masked self-attentional layers, which allows implicitly different weights to be assigned to different nodes in an adjacent set of nodes [24]. In this section, we would construct a GAT and take the network *HN* as its input to learn topological representations for nodes in *N* according to the following steps:

Step1 (Encoder): For any given node i in *HN*, let N_i denote the set of neighboring nodes of i in *N*, then, for any node $j \in N_i$, the GAT would first calculate the attention score α_{ij} between i and j according to the following formulae:

$$\alpha_{ij} = softmax(e_{ij}) = \frac{\exp(e_{ij})}{\sum_{k \in N_i} \exp(e_{ik})} \tag{10}$$

$$softmax(x) = \frac{1}{1 + e^{-x}} \tag{11}$$

$$e_{ij} = LeakyRelu(\alpha[Wh_i || Wh_j]) \tag{12}$$

$$LeakyRelu(x) = \begin{cases} xx > 0 \\ \mu x otherwise \end{cases} \tag{13}$$

Here, α represents the computational operation of self-attention, W is the matrix of trainable weights, h_i denotes the feature representation of the node i (i.e., the *i-th* row of *N*), μ is the hypermeter and || denotes the concatenation operation.

Subsequently, the GAT would multiply the attention score α_{ij} with the feature representation h_j of each node in N_i and sum all these products up as follows:

$$h_i = \sigma\left(\sum_{j \in N_i} \alpha_{ij} W h_j\right) \tag{14}$$

Here, σ denotes the activation function.

After above Encoder step, obviously, we could obtain a matrix $Z = \begin{bmatrix} Z^r \\ Z^m \end{bmatrix} \in R^{(n_r + n_m) * l}$, where Z^r and Z^m represent the low-dimensional topological representation of drug nodes and microbe nodes in *HN* respectively.

Step2(Decoder) Based on the matrix Z, it was easy to see that we could take its inner product as a decoder:

$$ZZ = sigmoid(Z \bullet Z^T) \tag{15}$$

$$sigmoid(x) = \frac{1}{1 + e^{-x}} \tag{16}$$

Step3(Optimization) Considering that the decoded result ZZ should be close to the original inputted matrix *N*, we adopted the MSE loss function to compute the mean of the sum of squares of the differences between ZZ and *N* as follows:

$$L_{MSE} = \frac{1}{n_r + n_m} \sum_{k=1}^{n_r+n_m} ||ZZ(k) - N(k)||^2 \tag{17}$$

where $ZZ(k)$ and $N(k)$ denote the k-th row of ZZ and N respectively.

Thereafter, based on the Eq. (16), we would adopt the Adam optimizer [25] to optimize the results of topological representations for nodes in *HN*.

Learning attribute representations for nodes in *HN*

In this section, in order to effectively capture local and global topological intrinsic characteristics of nodes in *HN*, we further implemented an improved random walk with restart (RWR) on S_r, where the RWR was defined as follows [26]:

$$q_i^{l+1} = \varphi M q_i^l + (1 - \varphi)\varepsilon_i \tag{18}$$

Here, φ is the restart probability and set as 0.1, M denotes the transition probability matrix, and $\varepsilon_i \in R^{1\times m}$ is the initial probability vector of the node i, which is defined as follows:

$$\varepsilon_{ij} = \begin{cases} 1 if i = j \\ 0 otherwise \end{cases} \tag{19}$$

Based on above RWR process, it was easy to know that we could obtain a novel matrix S_r^{mm}.

In addition, let n_d denote the number of newly downloaded diseases, similar to construction of the adjacency matrix A, we could obtain an adjacency matrix $D \in R^{n_r \times n_d}$ based on these newly-downloaded known drug–disease associations as well. And then, for any two given drug nodes i and j in *HN*, we could calculate the cosine similarity $S_r^{dis}(i,j)$ between them as follows:

$$S_r^{dis}(i,j) = cos(D(i), D(j)) = \frac{D(i) \cdot D(j)}{||D(i)|| \times ||D(j)||} \tag{20}$$

Here, $D(i)$ denotes the *i-th* row of D.

Moreover, in a similar way, we could further calculate the drug structural similarity matrix S_r^{che} based on the dataset downloaded from the SIMCOMP2 [27], which measured the drug similarity based on the drug chemical structure information.

Hence, through integrating all these matrices A, S_r^{mm}, S_r^{dis} and S_r^{che}, it is easy to see that we could obtain a novel drug attribute matrix A^r as follows:

$$A^r = [A; S_r^{che}; S_r^{mm}; S_r^{dis}] \tag{21}$$

Similarly, after applying the improved RWR on S_m, we could obtain a new matrix S_m^{rr}.

And in addition, through adopting the method proposed by Kamneva [28], which calculated the functional similarity between microbes based on a microbial protein–protein functional association network, we could obtain a new matrix S_m^f as well. Here, in the microbial protein–protein functional association network, the nodes represent any gene family encoded by the genome and the edges denote genetic neighbor scores based on STRING database. The functional similarity between microbes was

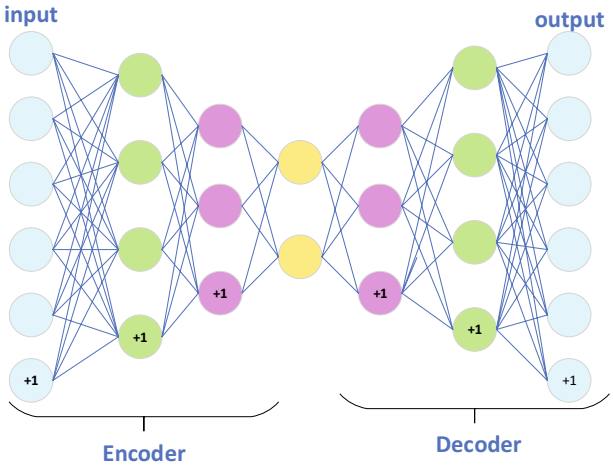

Fig. 2 The overall architecture of GSAMDA

calculated as the ratio of the link scores connecting the two microbes to the sum of all the link scores of the two microbial gene families.

Moreover, similar to the construction of S_r^{dis}, based on the dataset of newly downloaded known microbe–disease associations, for any two given microbe nodes i and j in HN, we could calculate the cosine similarity $S_m^{dis}(i,j)$ between them in a similar way as well.

Hence, through integrating all these matrices A^T, S_m^f, S_m^{rr} and S_m^{dis}, it is obvious that we could obtain a novel microbe attribute matrix A^m as follows:

$$A^m = [A^T; S_m^f; S_m^{rr}; S_m^{dis}] \tag{22}$$

Thereafter, after taking above two kinds of attribute matrices A^r and A^m as input of the SAE respectively, we could learn a unique attribute representation for each node in HN as well, where the structure of SAE was shown in the following Fig. 2.

From observing above Fig. 2, it is easy to see that the SAE consists of the following steps:

Step1(Encoder) Let h and x represent the hidden layer and the input layer of the SAE respectively, the encoder could be formulated as follows:

$$h_{W,b} = \sigma(W_{encoder}x(i) + b_{encoder}) \tag{23}$$

Step2(Decoder) The decoder adopted the same structure as the encoder, which was defined as follows:

$$y_{W,b} = \sigma(W_{decoder}h + b_{decoder}) \tag{24}$$

where W is the weight matrix between two layers and b is the bias term.

Moreover, in order to ensure the sparsity of the hidden layer, we would add a penalty term in the SAE as follows:

$$P_{penalty} = \sum_{t=1}^{S_2} KL(\rho||\widehat{\rho}_t) \tag{25}$$

where S_2 is the number of neurons in the hidden layer, $\widehat{\rho_t}$ represents the average activity of hidden neuron t, $KL(\rho||\widehat{\rho_t})$ is the relative entropy between two Bernoulli random variables with mean ρ and mean $\widehat{\rho_t}$ and is defined as follows:

$$KL(\rho||\widehat{\rho_t}) = \rho log\frac{\rho}{\widehat{\rho_t}} + (1-\rho)log\frac{1-\rho}{1-\widehat{\rho_t}} \tag{26}$$

Hence, after inputting the drug attribute matrix A^r and the microbe attribute matrix A^m into the SAE, we could obtain the output matrices A^{rr} and A^{mm} respectively.

Step3(Optimization) In the SAE, we adopted the MSE loss function and the Adam optimizer for optimization as well. During optimization, the sparse penalty term would be added to the loss function as follows (Taking the drug attribute matrix as an example):

$$L_{sparse} = \frac{1}{n_r}\sum_{k=1}^{n_r}||A^{rr}(k) - A^r(k)||^2 + \beta P_{penalty} \tag{27}$$

Here, β is the weight of the sparse penalty and will be set to 0.1. $A^{rr}(k)$ and $A^r(k)$ represent the *k-th* row of A^{rr} and A^r respectively.

After training the SAE, we could adopt the trained SAE to learn and obtain the low dimensional drug attribute representation matrix $\widetilde{A^r} \in R^{n_r * k}$ and microbe attribute representation matrix $\widetilde{A^m} \in R^{n_m * k}$ simultaneously.

Constructing feature matrices for microbes and drugs

Based on above drug topological representation matrix Z^r, drug attribute representation matrix $\widetilde{A^r}$, drug structural similarity matrix S_r^{che}, drug cosine similarity matrix S_r^{dis}, drug similarity matrix S_r^{mm} and the original adjacency matrix A, inspired by Xuan et al. [29], we could construct a novel drug feature matrix F_r as follows:

$$F_r = [Z^r; \widetilde{A^r}; S_r^{che}; A; S_r^{dis}; A; S_r^{mm}; A] \tag{28}$$

Similarly, based on above microbe topological representation matrix Z^m, microbe attribute representation matrix $\widetilde{A^m}$, microbe functional similarity matrix S_m^f, microbe cosine similarity matrix S_m^{dis}, microbe similarity matrix S_m^{rr} and the original adjacency matrix A^T, we can construct a novel microbe feature matrix F_m as follows:

$$F_m = [Z^m; \widetilde{A^m}; A^T; S_m^{fun}; A^T; S_m^{dis}; A^T; S_m^{rr}] \tag{29}$$

Computing predicted scores for microbe–drug pairs

The multiplication of two vectors is an effective means of simulating the interaction, which emphasizes the commonality of the interaction and dilutes the difference information of the interaction. Hence, for any given drug r_i and microbe m_j, we could obtain the predicted score between them by calculating the inner product of their feature representations as follows:

$$S(r_i, m_j) = Sigmoid(F_r(r_i) \bullet F_m(m_j)^T) \tag{30}$$

Table 2 The AUCs, AUPRs and Accuracy of compared methods based on datasets MDAD and aBiofilm under fivefold CV

Methods	AUC		AUPR		Accuracy	
	MDAD	aBiofilm	MDAD	aBiofilm	MDAD	aBiofilm
HMDAKATZ	0.8712 ± 0.0010	0.8993 ± 0.0021	0.2327 ± 0.0068	0.3066 ± 0.0077	0.9774	0.9796
LAGCN	0.8533 ± 0.0070	0.8641 ± 0.0109	0.3571 ± 0.0051	0.3671 ± 0.0055	0.9413	0.9373
NTSHMDA	0.8483 ± 0.0020	0.8610 ± 0.0022	0.1892 ± 0.0056	0.1962 ± 0.0078	0.9896	**0.9882**
HMDA-Pred	0.7987 ± 0.0030	0.8053 ± 0.0040	0.0236 ± 0.0009	0.0284 ± 0.0006	0.9794	0.9806
BPNNHMDA	0.8410 ± 0.0320	0.8438 ± 0.0186	0.0319 ± 0.0105	0.0476 ± 0.0067	0.9894	0.9869
GSAMDA	**0.9496 ± 0.0005**	**0.9308 ± 0.0120**	**0.4436 ± 0.0007**	**0.4510 ± 0.0051**	**0.9896**	0.9880

Here, $F_r(r_i)$ denotes the i-th row of F_r and $F_m(m_j)$ denotes the j-th row of F_m.

Results

In this section, we first compared GSAMDA with five state-of-the-art competitive predictive methods based on databases MDAD and aBiofilm separately. And then, we conducted the hyperparameter sensitivity analysis to decide the best parameters. Finally, we implemented case studies on two selected drugs and two selected microbes to further demonstrate the performance of GSAMDA.

Comparison with state-of-the-art methods

As predicting microbe–drug associations is a new problem, there are few computational methods and codes available, therefore, we would compare our method GSAMDA with some representative methods for link prediction problems in this section. Among them, HMDAKATZ [14] is a KATZ-based method proposed for microbe–drug associations prediction. LAGCN [30] is a graph convolutional network with attention mechanism based method designed to infer potential drug–disease associations. NTSHMDA [31] is a model based on random walk with restart for microbe–disease associations prediction. HMDA-Pred [32] integrated multiple data types and adopted the Network Consistency Projection (NCP) technique to detect latent microbe–disease associations. BPNNHMDA [33] designed a novel neural network to infer microbe–disease associations.

During experiments, we settled with the original parameters for all these competitive methods and ran them on the MDAD and aBiofilm datasets respectively for a fair comparison. In addition, we adopted the framework of fivefold cross validation (CV) in Cai et al. [34] to evaluate these methods, in which, we randomly selected 20% of known associations and 20% of unknown associations as the testing set, and the remaining 80% of known associations and unknown associations as the training set. We run the fivefold CV for 10 times and the AUROCs, AUPR and the best Accuracy of all compared methods were shown in Table 2. The best ROC curves and PR curves of these six competitive methods based on datasets MDAD and aBiofilm were drawn in Figs. 3 and 4, respectively.

The indicators including true positive rate(TPR), false positive rate(FPR), precision and recall related to ROC curve and PR curve were calculated as follows:

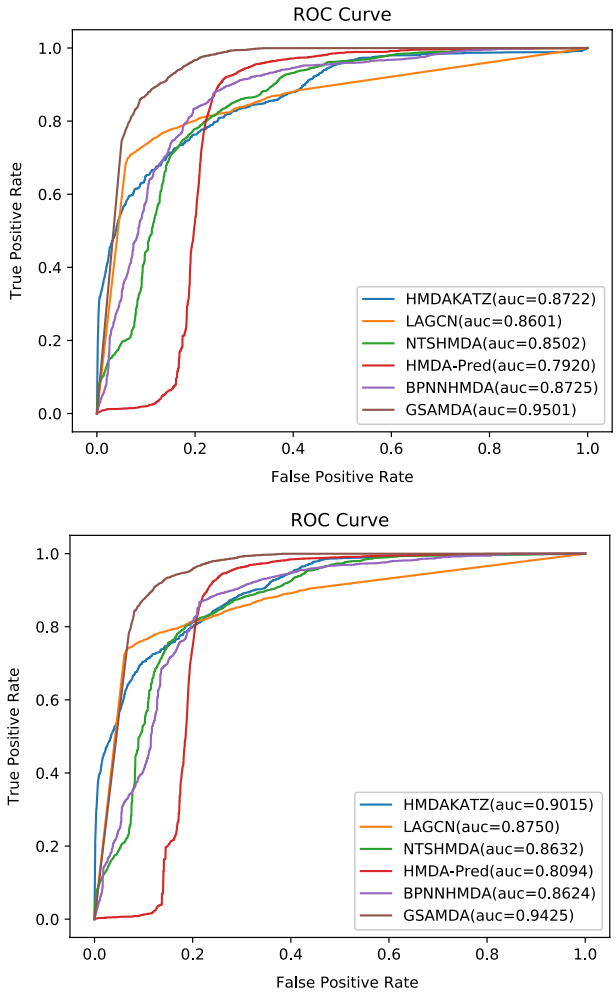

Fig. 3 **a** ROC curves of six competitive methods based on the MDAD dataset. **b** ROC curves of six competitive methods based on the aBiofilm dataset

$$TPR = \frac{TP}{TP + FN} \tag{31}$$

$$FPR = \frac{FP}{TN + FP} \tag{32}$$

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

$$Recall = \frac{TP}{TP + FN} \tag{34}$$

In addition, the accuracy is defined as below:

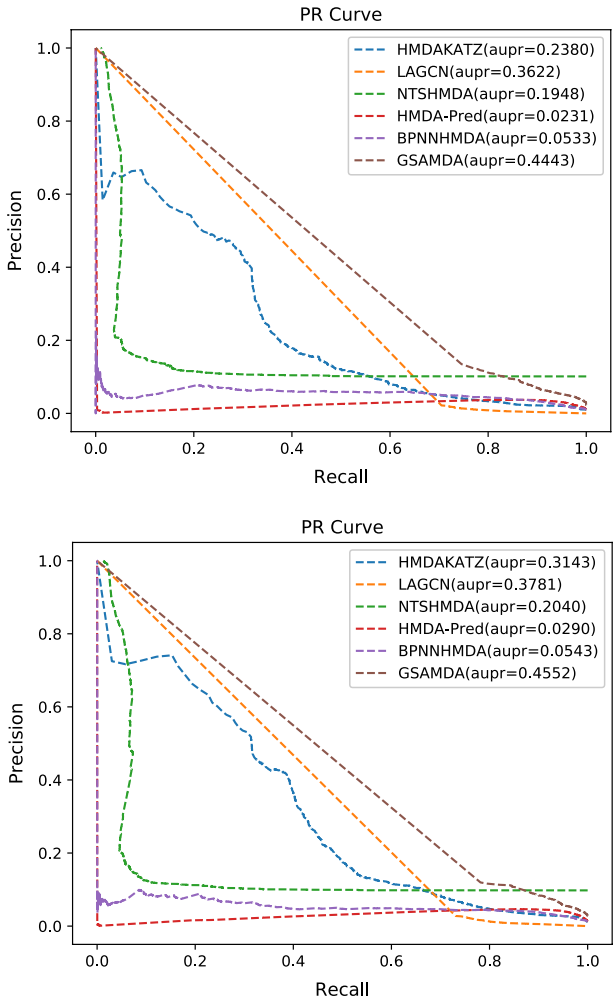

Fig. 4 **a** PR curves of six competitive methods based on the MDAD dataset. **b** PR curves of six competitive methods based on the aBiofilm dataset

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{35}$$

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

As shown in Table 2, it is obvious that GSAMDA can achieve the highest AUC values of 0.9496 ± 0.0005 and 0.9308 ± 0.0120 respectively based on both MDAD and aBiofilm, while HMDAKATZ can achieve the second highest AUC values of 0.8712 ± 0.0010 and 0.8993 ± 0.0021separately based on both MDAD and aBiofilm, which are 8.19% and 4.39% lower than that of GSAMDA respectively. Meanwhile, GSAMDA also obtained the highest AUPR values of 0.4436 ± 0.0007 and 0.4510 ± 0.0051 respectively based on both MDAD and aBiofilm. Moreover, the best accuracy of GSAMDA performs better than most compared methods.

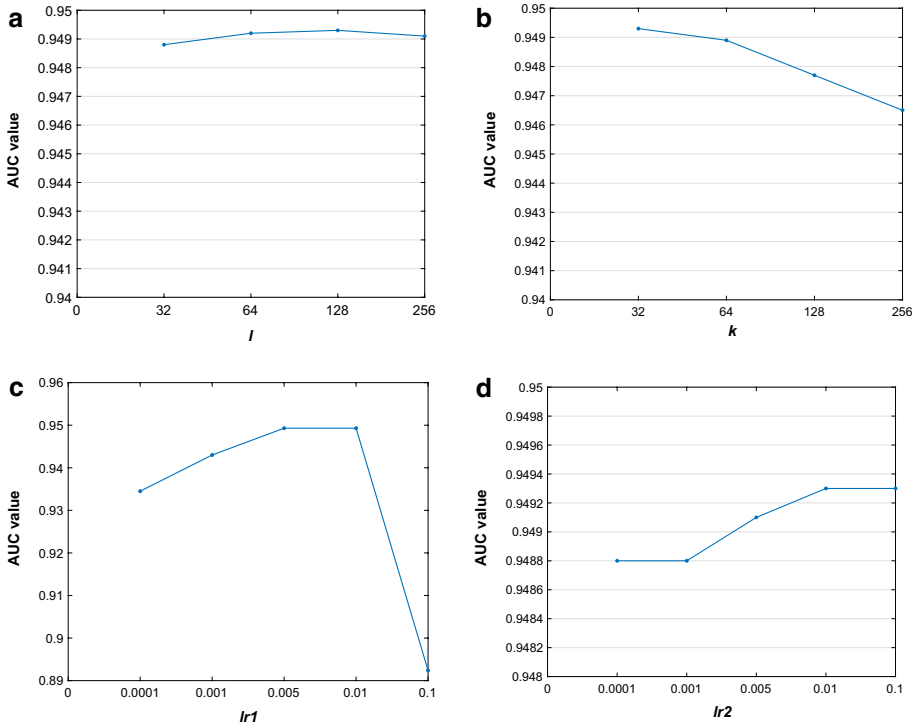

Fig. 5 Analysis of the impact of hyperparameters on performance of GSAMDA. The subfigures from (**a**) to (**d**) show the AUC values of the related values of the dimension of node topological representation and node attribute representation, the learning rate of GAE and SAE, respectively

Hyperparameter sensitivity analysis

Considering that there are several hyperparameters in our model GSAMDA including the dimension of node topological representation l, the dimension of node attribute representation k, and the learning rate lr_1 and lr_2 in GAE and SAE separately, 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 these parameter values.

First, we tested the dimension l and k in the range of {32, 64, 128, 256}, and illustrated the experimental results in Fig. 5a and b, respectively, from which, we found that the dimension has a subtle impact on the performance of GSAMDA. As shown in Fig. 5a and b, when l was set to 128 and k was set to 32, the performance would be the best. Next, through experimental results, we found that these two parameters for learning rate were important for the performance of GSAMDA, too high or too low of their values would both cause performance degradation of GSAMDA. In experiments, we selected lr_1 and lr_2 in the range of {0.0001, 0.001, 0.005, 0.01, 0.1}, and showed the results in Fig. 5c and d separately, from which, it is easy to see that GSAMDA can obtain the highest AUC values while both lr_1 and lr_2 are set to 0.01.

Case study

To further validate the performance of GSAMDA, in this section, we would select two popular drugs, Ciprofloxacin and Moxifloxacin, and two microbes, Human immunodeficiency virus type 1 and Mycobacterium tuberculosis, for case studies. During

Table 3 The top 20 predicted Ciprofloxacin-associated microbes

Microbe	Evidence	Microbe	Evidence
Bacillus Subtilis	PMID:33218776	Human Immunodeficiency Virus 1	PMID:9566552
Burkholderia Cenocepacia	PMID:27799222	Klebsiella Pneumoniae	PMID:27257956
Burkholderia Multivorans	PMID:19633000	Listeria Monocytogenes	PMID:28355096
Candida Albicans	PMID:31471074	Mycobacterium Tuberculosis	PMID:30020039
Actinomyces Oris	Unconfirmed	Pseudomonas Aeruginosa	PMID:33875431
Clostridium Perfringens	PMID:29978055	Salmonella Enterica	PMID:6933017
Enteric Bacteria and Other Eubacteria	PMID:27436461	Serratia Marcescens	PMID:23751969
Enterococcus Faecalis	PMID:27790716	Staphylococcus Aureus	PMID:32488138
Escherichia Coli	PMID:26607324	Staphylococcus Epidermidis	PMID:28481197
Haemophilus Influenzae	PMID:27292570	Staphylococcus Epidermis	PMID:10632381

The first column records top 10 microbes, while the third column records top 11–20 microbes

Table 4 The top 20 predicted Moxifloxacin-associated microbes

Microbe	Evidence	Microbe	Evidence
Bacillus Subtilis	PMID:30036828	Staphylococcus Aureus	PMID:31689174
Candida Albicans	PMID:21108571	Staphylococcus Epidermidis	PMID:31516359
Clostridium Perfringens	PMID:29486533	Staphylococcus Epidermis	PMID:11249827
Enteric Bacteria and Other Eubacteria	Unconfirmed	Stenotrophomonas Maltophilia	PMID:27257956
Enterococcus faecalis	PMID:31763048	Streptococcus Mutans	PMID:29160117
Escherichia Coli	PMID:31542319	Streptococcus Pneumoniae	PMID:31542319
Haemophilus Influenzae	PMID:11856249	Vibrio Harveyi	Unconfirmed
Listeria Monocytogenes	PMID:28739228	Burkholderia Cenocepacia	Unconfirmed
Pseudomonas Aeruginosa	PMID:31691651	Human Immunodeficiency Virus 1	PMID:18441333
Salmonella Enterica	PMID:22151215	Actinomyces Oris	PMID:26538502

The first column records top 10 microbes, while the third column records top 11–20 microbes

experiments, we selected the top 20 microbes or drugs predicted by GSAMDA based on MDAD for investigation, and then verified that whether these top 20 predicted microbes or drugs have been reported by PubMed literatures.

Ciprofloxacin is a fluorinated quinolone antibiotic with high activity against a wide spectrum of gram-positive and gram-negative bacteria, including methicillin-resistant Staphylococcus aureus, Enterobacteriaceae, and Pseudomonas aeruginosa [35]. For example, *Mycobacterium avium* is highly susceptible to Ciprofloxacin [36]. And it is validated that Ciprofloxacin is an active agent against *Candida albicans* [37]. Besides, the Moxifloxacin [38] is a fluoroquinolone antibiotic, which can treat the social acquired pneumonia caused by Staphylococcus aureus, influenza bacillus, pneumococcus, acute attack of chronic bronchitis, acute sinusitis and so on. Gislason et al. revealed a two-component system that sensitized Burkholderia cenocepacia to moxifloxacin after depletion of the response regulator [39]. Tahoun et al. found that Listeria monocytogenes' antimicrobial susceptibility was most frequently observed for moxifloxacin [40]. Chon et al. demonstrated that most isolates of Clostridium perfringens were susceptible to moxifloxacin [41]. As shown in Tables 3 and 4, among these top 20 predicted ciprofloxacin-associated and moxifloxacin-associated microbes, we found 19 and 17 microbes having been reported by PubMed

Table 5 The top 20 predicted Mycobacterium tuberculosis-associated drugs

Drug	Evidence	Drug	Evidence
Ciprofloxacin	PMID:16270765	Viomycin	PMID:16048924
Epigallocatechin Gallate	PMID:17996734	Capreomycin	PMID:29311078
Tobramycin	PMID:19723387	Ethambutol	PMID:27806932
Curcumin	PMID:24631908	Cloxacillin	PMID:25104892
Vancomycin	PMID:33508482	Aminosalicylic Acid	PMID:26033719
LL-37	PMID:26351280	Amikacin	PMID:29311078
Triclosan	PMID:19130456	3-(pyridin-3-yl)-5-(3-ethoxy-4-hydroxybenzylidene)-2-thioxo-thiazolidin-4-one	Unconfirmed
Ceftazidime	PMID:28875168	BMAP-28	Unconfirmed
Farnesol	Unconfirmed	3-(4-fluorophenyl)-5-(3-(allyloxy)-4-hydroxybenzylidene)-2-thioxothi-azolidin-4-one	Unconfirmed
Pyrazinamide	PMID:26521205	Azithromycin	PMID:7849341

The first column records top 10 drugs, while the third column records top 11–20 drugs

Table 6 The top 20 predicted Human immunodeficiency virus type 1-associated drugs

Drug	Evidence	Drug	Evidence
Abacavir	PMID:11797183	Didanosine	PMID:9107385
Amprenavir	PMID:10868554	Efavirenz	PMID:10952598
Atovaquone	PMID:8780816	Emtricitabine	PMID:31879782
Cefditoren	Unconfirmed	Farnesol	Unconfirmed
Cefixime	Unconfirmed	Fosamprenavir	PMID:19515730
Ceftazidime	PMID:11527042	Ganciclovir	PMID:1510405
Cidofovir	PMID:10926733	Indinavir	PMID:8970946
Ciprofloxacin	PMID:9566552	Lamivudine	PMID:12543687
Darunavir	PMID:31879782	LL-37	PMID:17627504
Delavirdine	PMID:9107385	Lopinavir	PMID:20836579

The first column records top 10 drugs, while the third column records top 11–20 drugs

literatures, which means that the prediction performance of GSAMDA can reach up to 95% and 85%, and demonstrates as well that GSAMDA can achieve satisfactory performance.

With regards to microbes, mycobacterium tuberculosis is a species of gram-positive, aerobic bacteria that is the etiological agent of tuberculosis which is the leading cause of death from bacterial infections [42]. The bacteria can infect various organs in the human body, causing pulmonary tuberculosis the most common. Moreover, human immunodeficiency virus type 1(HIV-1) is a member of the lentivirus ('slow-acting') genus of the family Retroviridae [43]. It is the cause of the Acquired Immunodeficiency Syndrome (AIDS) which is a deadly infectious disease. The top 20 predicted mycobacterium tuberculosis-associated and human immunodeficiency virus type 1-associated drugs are shown in Tables 5 and 6, respectively, from which, we can see that there are 16 and 17 out of top 20 predicted drugs having been validated by PubMed literatures, which further demonstrates the predictive power of GSAMDA as well.

Discussion and conclusion

Increasing researches have shown that microbes are closely related to human health. Predicting microbe–drug associations can promote microbe-derived therapy and drug discovery. However, traditional wet experiments are time-consuming and expensive and few predictive computational models for microbe–drug associations have been studied. An effective predictive computational model will be a great help for microbe–drug associations discovery.

In this paper, we designed a novel calculation model called GSAMDA based on both GAT and SAE for possible microbe–drug association prediction. In GSAMDA, we first constructed a heterogeneous network based on known microbe–drug associations. And then, the GAT- and SAE-based modules were established to learn the topological representations and the attribute representations of microbe and drug nodes in the heterogeneous network respectively. Finally, through combining the node topological representations and attribute representations with multiple original node features of nodes in the heterogeneous network, the microbe feature matrix and drug feature matrix would be constructed to infer potential microbe–drug associations. Experimental results of both comparison with five state-of-the-art competitive prediction methods and case studies demonstrated the superior performance of GSAMDA and its great potential for drug discovery.

It should be noted that some limitations still exist in GSAMDA. Firstly, the microbe–drug association matrix is sparse and it will affect the performance of the model to some extent. Moreover, not all microbes(drugs) have diseases associated with them, and there are some defects in using microbe(drug)–disease association as attribute feature. Finally, we can incorporate more biological information, such as microbe-microbe interactions and drug–drug interactions, to improve the performance of the model.

Acknowledgements
Not applicable.

Author contributions
YQT and LW designed the model and conducted the experiments, YQT and XYW wrote this paper. LW, JZ, LAK, BZ and ZZ provide suggestions and revised the manuscript. All authors read and approved the final manuscript.

Funding
The National Natural Science Foundation of China (No.62272064, No.61873221) and the Key project of 321 Changsha Science and technology Plan (No. KQ2203001).

Availability of data and materials
The data and code can be found online at: https://github.com/tyqGitHub/TYQ.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

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

Received: 12 October 2022 Accepted: 14 November 2022
Published online: 18 November 2022

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, Bäckhed 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. ElRakaiby M, Dutilh BE, Rizkallah MR, et al. Pharmacomicrobiomics: the impact of human microbiome variations on systems pharmacology and personalized therapeutics. OMICS. 2014;18(7):402–14.
6. Ley RE, Turnbaugh PJ, Klein S, et al. Human gut microbes associated with obesity. Nature. 2006;444(7122):1022–3.
7. Durack J, Lynch SV. The gut microbiome: relationships with disease and opportunities for therapy. J Exp Med. 2018;216(1):20–40.
8. Schwabe RF, Jobin C. The microbiome and cancer. Nat Rev Cancer. 2013;13(11):800–12.
9. Xiang Y-T, Li W, Zhang Q, et al. Timely research papers about COVID-19 in China. Lancet. 2020;395(10225):684–5.
10. McCoubrey LE, Gaisford S, Orlu M, et al. Predicting drug–microbiome interactions with machine learning. Biotechnol Adv. 2022;54: 107797.
11. 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.
12. Sun Y-Z, Zhang D-H, et al. MDAD: a special resource for microbe–drug associations. Front Cell Infect Microbiol. 2018;8:424.
13. 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.
14. 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
15. 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.
16. 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.
17. Long Y, Wu M, Kwoh CK, et al. Predicting human microbe–drug associations via graph convolutional network with conditional random field. Bioinformatics. 2020;36(19):4918–27.
18. 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.
19. 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.
20. Dayun L, Junyi L, Yi L, et al. MGATMDA: predicting microbe–disease associations via multi-component graph attention network. IEEE/ACM Trans Comput Biol Bioinform. 2021. https://doi.org/10.1109/TCBB.2021.3116318.
21. Jiang HJ, Huang YA, You ZH. SAEROF: an ensemble approach for large-scale drug–disease association prediction by incorporating rotation forest and sparse autoencoder deep neural network. Sci Rep. 2020;10(1):1–11.
22. 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. Briefings Bioinform. 2022;23(3):bbac080.
23. Xu D, Xu H, Zhang Y, et al. MDAKRLS: predicting human microbe–disease association based on Kronecker regularized least squares and similarities. J Transl Med. 2021;19(1):1–12.
24. Veličković P, Cucurull G, Casanova A, et al. Graph attention networks. arXiv preprint arXiv:1710.10903, 2017.
25. Kingma D, Ba J. Adam: a method for stochastic optimization. Comput Sci. 2014;10(22):1–15.
26. Köhler S, Bauer S, Horn D, et al. Walking the interactome for prioritization of candidate disease genes. Am J Hum Genet. 2008;82(4):949–58.
27. Hattori M, Tanaka N, Kanehisa M, et al. SIMCOMP/SUBCOMP: chemical structure search servers for network analyses. Nucleic Acids Res. 2010;38(Suppl2):W652–6.
28. Kamneva OK. Genome composition and phylogeny of microbes predict their co-occurrence in the environment. PLoS Comput Biol. 2017;13(2): e1005366.
29. Xuan P, Gao L, Sheng N, et al. Graph convolutional autoencoder and fully-connected autoencoder with attention mechanism based method for predicting drug–disease associations. IEEE J Biomed Health Inform. 2020;25(5):1793–804.
30. Yu Z, Huang F, Zhao X, et al. Predicting drug–disease associations through layer attention graph convolutional network. Briefings Bioinform. 2020;22(4):bbaa243.
31. 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.
32. Fan Y, Chen M, Zhu Q, et al. Inferring disease-associated microbes based on multi-data integration and network consistency projection. Front Bioeng Biotechnol. 2020;8:831.
33. Li H, Wang Y, Zhang Z, Tan Y, Chen Z, Wang X, Pei T, Wang L. BPNNHMDA: identifying microbe-disease associations based on a novel back propagation neural network model. IEEE/ACM Trans Comput Biol Bioinform. 2021;18(6):2502–13.
34. Cai L, Lu C, Xu J, et al. Drug repositioning based on the heterogeneous information fusion graph convolutional network. Brief Bioinform. 2021;22(6):bbab319.
35. Terp DK, Rybak MJ. Ciprofloxacin. Drug Intell Clin Pharm. 1988;35(4):373–447.
36. Cho EH, Huh HJ, Song DJ, et al. Differences in drug susceptibility pattern between *Mycobacterium avium* and *Mycobacterium intracellulare* isolated in respiratory specimens. J Infect Chemother Off Jo Japan Soc Chemother. 2017;24(4):315–8.
37. Hacioglu M, Haciosmanoglu E, Birteksoz-Tan AS, et al. Effects of ceragenins and conventional antimicrobials on *Candida albicans* and Staphylococcus aureus mono and multispecies biofilms. Diagn Microbiol Infect Dis. 2019;95(3):114863.
38. Barman Balfour JA, et al. Moxifloxacin. Drugs. 1999;59(1):115–39.

39. Gislason AS, Choy M, et al. Competitive growth enhances conditional growth mutant sensitivity to antibiotics and exposes a two-component system as an emerging antibacterial target in *Burkholderia cenocepacia*. Antimicrob Agents Chemother. 2017;61(1):00790.
40. Tahoun A, Elez R, Abdelfatah EN, et al. Listeria monocytogenes in raw milk, milking equipment and dairy workers: molecular characterization and antimicrobial resistance patterns. J Glob Antimicrob Resist. 2017;10:264–70.
41. Chon J-W, Seo K-H, Bae D, et al. Prevalence, toxin gene profile, antibiotic resistance, and molecular characterization of Clostridium perfringens from diarrheic and non-diarrheic dogs in Korea. JVS. 2018;19(3):368–74.
42. Koch A, Mizrahi V. Mycobacterium tuberculosis. Trends Microbiol. 2018;26(6):555–6.
43. Spector SA. Human immunodeficiency virus type-1. Ref Module Biomed Sci. 2014;11(28):1–12.

Publisher's Note