RESEARCH ARTICLE

DWPREPHI: a novel deep learning-based computational model to predict phage-host interaction *via* complex multi-dimensional biological information

Jiaye Li[1, 2], Hongxiang Xiao[1, 2, *]

[1]School of Information Science and Engineering, [2]Guangxi Key Laboratory of Embedded Technology and Intelligent System, Guilin University of Technology, Guilin, Guangxi, China.

Many studies project that, due to antibiotic misuse, phage therapy has been considered as one of the most promising alternatives for the treatment of human diseases infected by antibiotic-resistant bacteria. The identification of phage-host interactions (PHI) helps to explore the mechanisms by which bacteria respond to phages and provides new insights into effective therapeutic approaches. Computational models for predicting PHI are not only time/cost saving, but also more efficient and economical than traditional wet experiments. In this work, we proposed a deep learning based computational model named DWPREPHI to predict PHI through the combining DNA and protein sequence information. More specially, DWPREPHI first extracted information about the node properties of the interaction network by a natural language processing algorithm that initialized the node representations of the phage and the target bacterial host. The graph embedding algorithm, Deepwalk, was then used to extract link behavior information from the PHI network, and finally a deep neural network was applied to accurately detect interactions between phages and their bacterial hosts. On the drug-resistant bacteria dataset ESKAPE, DWPREPHI achieved a prediction accuracy of 92.25% and an AUC value of 0.9674 under the 5-fold cross-validation method, which was significantly better than other methods. In addition, three case studies were conducted for *E. coli*, *Pseudomonas aeruginosa*, and *Salmonella enterica* to further demonstrate the utility of the proposed model. Among the top 10 phages associated with these hosts, 7, 8, and 8 have been reported. These excellent experimental results suggested that the DWPREPHI model could provide reasonable candidates for sensitive bacteria for biological experiments in phage therapy.

Keywords: phage-host interactions; sequence information; graph embedding algorithm; deep neural network.

Abbreviations: phage-host interactions (PHI); area under roc curves (AUC); deep neural network (DNN); developed deep neural networks (DNNs); receptor-binding protein (RBP); receiver operating characteristic curve (roc); 5-fold cross-validation method (5-fold CV); random forest (RF); support vector machine (SVM); K Nearest Neighbors (KNN); Gradient Boosting Decision Tree (GBDT); Structural Deep Network Embedding (SDNE); Laplacian Eigenmaps (Lap).

[*]**Corresponding author:** Hongxiang Xiao, School of Information Science and Engineering, Guilin University of Technology, Guilin, Guangxi, China. Email: xhx@glut.edu.cn.

Introduction

Available research suggests that bacterial infections may be involved in the growth and development of a variety of diseases, including cholera [1], inflammatory bowel disease [2], colon cancer [3], tetanus [4], and different types of cancers [5-7]. Researchers discovered antibiotics in 1928 and have since used them extensively in clinical practice to treat serious bacterial diseases [8], saving countless lives. Unfortunately, due to the overuse of antibiotics, bacteria have developed resistance mechanisms [9]. In 2019, centers for disease control and Prevention reported that approximately 2.8 million cases of antibiotic-resistant infections

occur each year in the United States [10], resulting in more than 35,000 deaths; Similarly, in Europe, 33,000 people die each year from antibiotic-resistant infections [11]. Thus, there is an urgent need to develop new antibiotics or alternative therapies to avoid further deterioration of antibiotic-resistant infections. However, many pharmaceutical companies do not continue to develop new antibiotics due to high production costs, expected benefits and long development times [12]. Therefore, researchers are looking for alternative therapies to reduce antibiotic-resistant infections and to treat bacterial diseases. The ability of phages not only to destroy specific bacterial hosts but also to replicate exponentially has made phage therapy one of the most promising therapies for the treatment of bacterial diseases and antibiotic-resistant infections [13]. Predicting phage-host interactions (PHI) can help to understand whether phages can be used to treat bacterial diseases [14]. However, experimental validation methods for PHI require considerable time, human and financial resources. Thus, researchers have sought to develop computationally based methods for PHI to predict and screen target phages for the treatment of bacterial diseases to reduce the time and money costs required.

Studies have shown that proteins play a fundamental role in the biology of phages and hosts; thus, researchers have proposed methods to predict PHI based on protein sequences. For example, Leite *et al*. [15, 16] used primary structure sequences of phage and host proteins and classical machine learning classifiers, including RF, SVM, LR, k-nearest neighbor KNN, and multi-layer perceptron (MLP), to predict PHI. Zhu *et al*. [17] proposed a novel deep learning-based model named PHIHNE that predicts the phage-host interactions through heterogeneous network embedding methods. Zhou *et al*. [18] developed PHISDetector, which is used to predict phage–host interaction signals through machine learning based model. Galiez *et al*. [19] presented WIsH model, which performed a suited probabilistic approach to calculate the k-mer frequencies for host prediction. Although the existing methods have achieved good results in PHI prediction, there are still some limitations. First, there are thousands of experimentally validated PHI pairs in the database, but only a few hundred non-redundant PHI pairs are available for building prediction models [20]. This limitation hinders the development of high-performance predictive models. Second, most of the existing methods use phage and host DNA sequences or protein sequences to construct predictive models, but few of them are able to combine both types of sequence information [21]. Third, although prediction models have been built using various feature and machine learning techniques, these models often lack sufficient interpretability, hindering the elaboration of PHI prediction mechanisms [22]. In recent years, graph embedding algorithms have received much attention in the fields of cell biology and bioinformatics. Researchers have gradually started to experiment with applying such techniques to tackle different prediction tasks. As a typical model of graph neural networks, the Deepwalk algorithm uses a random walk-based approach to learn the topological features of the network and has recently been widely applied in the field of bioinformatics [23-25]. For example, Li *et al*. [26] proposed a method based on Deepwalk and network consistency projection for predicting circRNA-disease associations. Deepwalk was used in this work to learn features of the circRNA-disease association network and combine it with circRNA-circRNA, disease-disease similarity for predicting circRNA-disease potential correlations. In addition, some researchers have developed deep neural networks (DNNs) to improve the interpretability of predictive models [27]. Further developments in these techniques provide new perspectives on prediction accuracy [28].

In this paper, we proposed a novel PHI prediction model named DWPREPHI. The approach was based on the powerful graph embedding algorithm Deep Wander and the natural language processing algorithm Word2vec to solve various problems of PHI prediction.

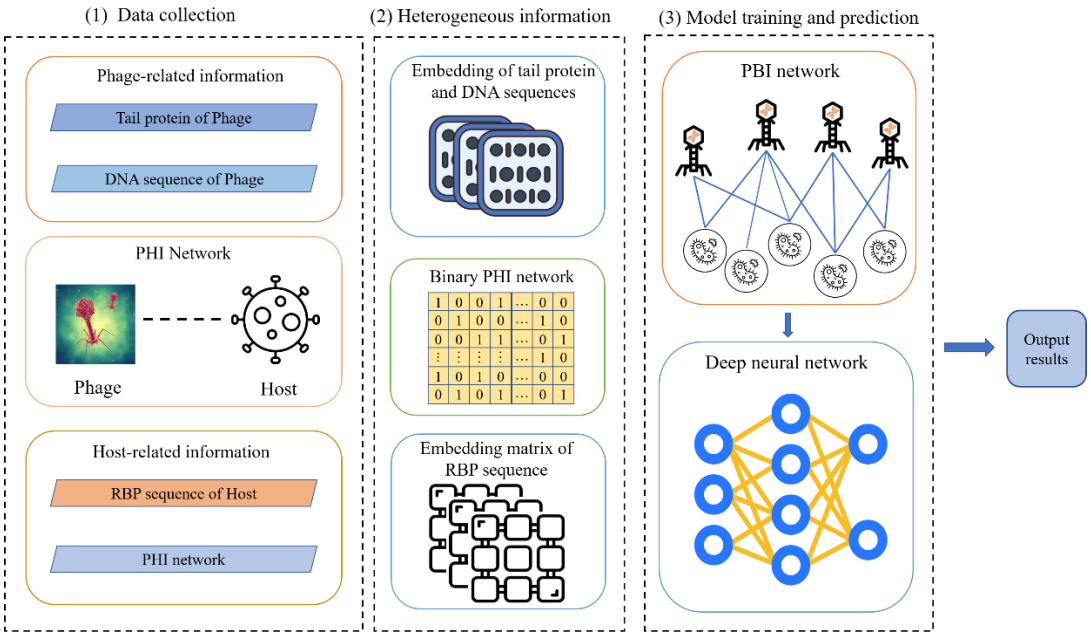

Figure 1. Workflow diagram of the DWPREPH prediction model proposed in this paper.

Specifically, we first constructed a phage-host interaction network to generalize the connection between phages and bacteria. The nodes in the network graph represented the phage and the target host, and the links between the nodes represented their interactions. Behavioral features were then captured from their interaction links using Deepwalk. A natural language processing algorithm, word2vec, was also used to encode the tail protein and DNA sequences of the phage and the receptor-binding protein (RBP) on the surface of the host to extract the attribute information. Finally, DWPREPHI integrated the behavioral and attribute information into a fusion matrix and then used a deep neural network (DNN) to achieve predictive classification. Comparison results with state-of-the-art machine learning classifiers as well as graph embedding methods demonstrated the feasibility and efficiency of the proposed model. A case study of three highly pathogenic bacteria further demonstrated the usefulness of the proposed model. The combined experimental results showed that the DWPREPHI model was well suited for predicting phage-host interactions. In future work, we hope that it will become a useful complementary tool for biology. The workflow diagram of the DWPREPH prediction model was shown in Figure 1.

Materials and Methods

Dataset description

The tail protein of the phage and the receptor binding protein on the surface of the host determines whether the phage can attach to the host. Also, a fundamental function of phage DNA is to direct the synthesis of its endogenous counterpart (tail protein). Therefore, we took these three key factors into account when constructing our predictive DWPREPHI model. In our experiments, we collected 1,170 DNA and protein information related to the tail structure of the target phage from three different public databases, including UniprotKB [29], UniRef [30], and Millard Lab (http://millardlab.org), together with information on the RBP sequences of their corresponding hosts. The dataset was dominated by ESKAPE (*Enterococcus faecalis*, *Acinetobacter baumannii*, *Pseudomonas aeruginosa*, *Staphylococcus aureus*, *Klebsiella pneumoniae*,

and *Enterobacter spp.*) pathogens [31], supplemented by *Escherichia coli*, *Salmonella enterica*, and *Clostridium difficile*. To reduce the computational load, identical sequences were removed [32]. We ended up with a collection of 1,232 phage-host pairs consisting of nine bacterial species.

Constructing the behavior features of PBI networks

Deepwalk, a widely used graph embedding method, was applied in this paper to extract behavior information from the links in the PBI networks [33]. The Deepwalk algorithm consists of two main kinds, the random walk algorithm and the Skip-gram algorithm. Specially, it applied the random walk of t length in each node v_i, and then utilized Skip-gram model to learn the embedding vectors from these nodes. The Skip-gram model could calculate the like hood of the length in window w and its object function was as follows:

$$\min_{\Phi} -\log \Pr\left(\left\{v_{i-w},...,v_{i+w}\right\} \setminus v_i \middle| \Phi\left(v_i\right)\right) \quad (1)$$

where Φ was a $(m\times n)\times d$ low-dimensional spaces matrix and denoted the representation of v_i. d represented the embedding dimension. The Skip-Gram framework further approximated the above conditional probabilities using the following assumptions:

$$Pr\left(\left\{v_{i-w},...,v_{i+w}\right\} \setminus v_i \middle| \Phi\left(v_i\right)\right) = \prod_{j=i-w, j\neq i}^{i+w} Pr\left(v_j \middle| \Phi\left(v_i\right)\right) \quad (2)$$

To reduce the computing time of $Pr\left(v_j \middle| \Phi\left(v_i\right)\right)$, the *softmax* function of the hierarchy was used to decompose conditional probabilities by assigning the vertices of a walking sequence to the leaves of a binary tree, which was shown in equation (3).

$$Pr\left(v_j \middle| \Phi\left(v_i\right)\right) = \prod_{l=1}^{\lceil \log|V| \rceil} 1/\left(1+e^{-\Phi(v_i)g\Psi(b_t)}\right) \quad (3)$$

where $\left(b_0, b_1, ..., b_{\log|v|}\right)$ represented the node sequence from start node to finish node v_j. $\Psi\left(b_l\right)$ was the corresponding embedding vectors about b_l.

After performing the Deepwalk algorithm in PHI network, we obtained the embedding matrix Φ. Each row of Φ represented to a d-dimensional embedding vector of potential topological representations of each node. Thus, the cosine similarity between the two embedding vectors could be calculated as the similarity between these nodes. The similarity formula of phages and host nodes was shown as follows:

$$Sim\left(v_i, v_j\right) = \frac{\sum_{k=1}^{d} \Phi\left(v_i, k\right)\Phi\left(v_j, k\right)}{\sqrt{\sum_{k=1}^{d} \Phi\left(v_i, k\right)^2}\sqrt{\sum_{k=1}^{d} \Phi\left(v_j, k\right)^2}} \quad (4)$$

where $\Phi\left(v_i, k\right)$ and $\Phi\left(v_j, k\right)$ represented the *k*-th component of the embedding vector $\Phi\left(v_i\right)$ and $\Phi\left(v_j\right)$. According to formula (4), we ccould construct the phage topological similarity matrix Sim_p and host topological similarity matrix Sim_h.

Constructing the attribute features of PHI network

In the DWPREPHI model, a Word2vec algorithm [34] based on natural language processing techniques was used to encode the DNA and tail proteins of the phage and the receptor binding proteins of the host, thereby capturing information about the node properties of the phage-host interaction network. The difference between CBOW and Skip-gram models [35] is that CBOW uses the context to predict the current word, while Skip-gram uses the current word to predict the context. Skip-gram is more efficient if the training data is not very large. In our experiments, given the size of the PHI dataset, we chose the CBOW model to learn more words,

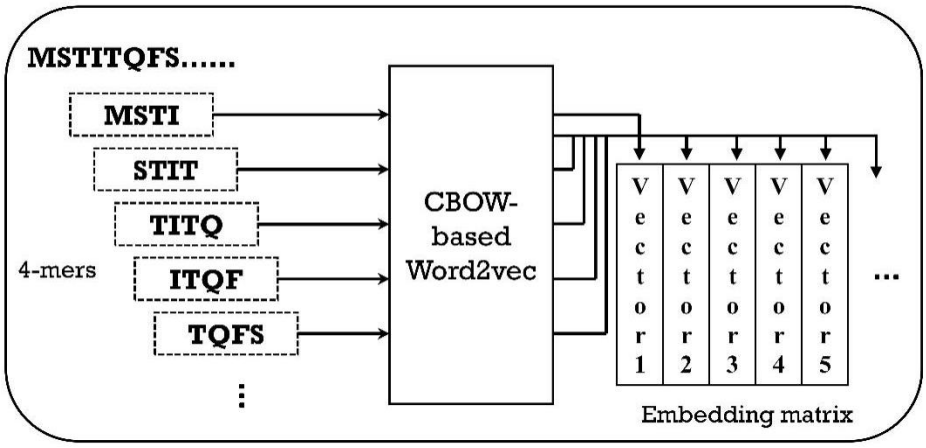

Figure 2. Flowchart of the word2vec algorithm based on the CBOW framework.

thus speeding up the training time. The Word2vec algorithm was used to encode a matrix of DNA and protein sequences to extract information about the node properties of the topological network. The method involved representing the sequences as multiple K-mers words [36]. Assuming a receptor protein sequence MSTITQF is given a 4-mer length, it can be divided into several words such as MSTI, STIT, TITQ, and ITQF. To speed up the training, we chose the Word2vec algorithm based on the continuous bag-of-words model to extract the word vectors. Here, DNA sequences, protein sequences, and k-mers corresponded to sentences and words in natural language, respectively. In this work, the trained Word2vec model would generate a 64-dimensional embedded word vector as a means to extract the attribute features of the network nodes. In previous studies, 4-mer had been shown to be the length at which optimal prediction accuracy was obtained in a five-fold cross-validation framework. The flowchart of the Word2vec algorithm employed in this paper was shown in Figure 2.

Deep neural network

Artificial neural networks are originally inspired by neural networks in the brain and consist of multiple layers of interconnected computational units (neurons). The depth of a neural network corresponds to the number of hidden layers, while the width corresponds to the maximum number of neurons in it. Artificial neural networks with a multilayer structure (two or more hidden layers) are called deep neural networks [37]. In terms of its structure, a DNN is a multi-layer stack of common modules. Features are first received at the input layer and then transformed non-linearly between multiple hidden layers. The average gradient is calculated to adjust the design weights accordingly before producing the final output. In addition, all neurons of the first hidden layer are connected to all neurons forming the input layer, while all neurons of the last hidden layer are connected to the output layer. The weighted sum of its inputs will then be calculated by the neurons and its output evaluated using a non-linear activation function. In this work, rectified linear units (*ReLU*) [38], tanh and *softmax* [39] were used as activation functions. More specifically, the tanh function was used in the input layer, while the activation functions in the hidden and output layers were the *ReLU* and *softmax* functions, respectively. A binary cross-entropy function [40] was used as the loss function. The Dropout learning algorithm [41] and the Adam optimiser [42] were also used to avoid overfitting and to speed up training. The entire network was defined as follows.

$$H_{i1}^{m} = \sigma_1\left(W_{i1}X_{i1} + b_{i1}\right),\ i = 1,\ldots,n \qquad (5)$$

$$H_{ij}^{m}=\sigma_1\left(W_{ij}H_{i(j-1)}+b_{ij}\right),\ j=1,\mathrm{L}\ ,n;\ j=2,\mathrm{L}\ ,h_1;\ m=1,2 \quad (6)$$

$$H_{ik}^{3}=\sigma_1\left(W_{ik}\left(H_{ih_1}^{1}\oplus H_{ih_1}^{2}\right)+b_{ik}\right),\ i=1,\mathrm{L}\ ,n;\ k=h_1+1 \quad (7)$$

$$Loss=-\frac{1}{N}\sum_{i=1}^{N}\left[y_i\ln\left(\sigma_2\left(W_{ih_2}H_{ih_2}+b_{ih_2}\right)\right)+(1-y_i)\ln\left(1-\sigma_2\left(W_{ih_2}H_{ih_2}+b_{ih_2}\right)\right)\right] \quad (8)$$

where *m* denoted the individual networks and *n* denoted the batch size of the PHI pairs used for network training. The depths of the fusion network and the two individual networks were denoted by h_2 and h_1 and represented the activation functions *ReLU* and *softmax* for the hidden and output layers, respectively. *x* and *H* corresponded to the batch training inputs and outputs of the layers. The variable *W* represented the weight matrix between the input, hidden, and output layers, and *b* was the bias term. In addition, $\oplus$ was the concatenation operator and *y* represented the corresponding desired output.

Performance evaluation indicators

In this study, the 5-fold cross-validation framework (5-fold CV) was used to compute a measure of the predictive performance of the DWPREPHI model [43]. We first divided the ESKAPE dataset into five random subsets of equal sample size, and then used four of these subsets as the training set and the remaining one as the test set. This process was repeated five times until each subset was used as the test set once and only once. Finally, the mean and standard deviation of these results were used as the predicted output of the model. In the experiments, Accuracy (ACC), Sensitivity (Sen), Specificity (Spec), Precision (Prec), and F1-score (F1) were used as the criteria for assessing the predictive ability of the DWPREPHI model. The corresponding formulae are shown below.

$$ACC.=\frac{TP+TN}{FP+TP+FN+TN} \quad (9)$$

$$Sen.=\frac{TP}{FN+TP} \quad (10)$$

$$Spec.=\frac{TN}{TN+FP} \quad (11)$$

$$Prec.=\frac{TP}{TP+FP} \quad (12)$$

$$F1=\frac{2\times Prec.\times Sen.}{Prec.+Sen.} \quad (13)$$

where TP, FP, TN, and FN represented true positive, false positive, true negative, and false negative, respectively. ROC curves and PR (precision-recall) curves were also plotted and the area under the ROC curve (AUC value) was calculated to numerically demonstrate the predictive performance of the proposed model.

Results and discussion

Evaluation of prediction performance

In our experiments, we used a 5-fold cross-validation method (5-fold CV) to evaluate the model capability. Table 1 summarized the results obtained by the DWPREPHI model on the ESKAPE dataset. The proposed model achieved 92.25% accuracy, 98.70% sensitivity, 85.80% specificity, 87.45% precision, and 92.73% F1 value. Figure 3 plotted the ROC curves and PR curves generated by the proposed model on the ESKAPE dataset. The DWPREPHI model generated AUC and PR values of 0.9674 and 0.9496. These experimental results indicated that the proposed model had good predictive performance and was effective in predicting the potential PHI pairs.

Comparison of the performance of different classifier models

To further validate the better performance of deep neural network-based classifiers for phage-host prediction, we compared the deep neural network module with a number of powerful machine learning classifiers. Specifically, the behavioral and attribute information extracted from the phage-host interaction network was kept unchanged and the DNN module was replaced with four popular classifier models,

Table 1. Predictive performance of the DWPREPHI model on the ESKAPE model.

5-fold	ACC. (%)	Sen. (%)	Spec. (%)	Prec. (%)	F1	AUC
Fold-1	93.00	98.78	87.22	88.55	93.38	0.9751
Fold-2	93.31	99.19	87.42	88.75	93.68	0.9791
Fold-3	90.37	97.77	82.96	85.16	91.03	0.9459
Fold-4	90.87	98.17	83.57	85.66	91.49	0.9599
Fold-5	93.71	99.59	87.83	89.11	94.06	0.9767
Average	92.25±1.52	98.70±0.74	85.80±2.33	87.45±1.88	92.73±1.37	0.9672±0.0142

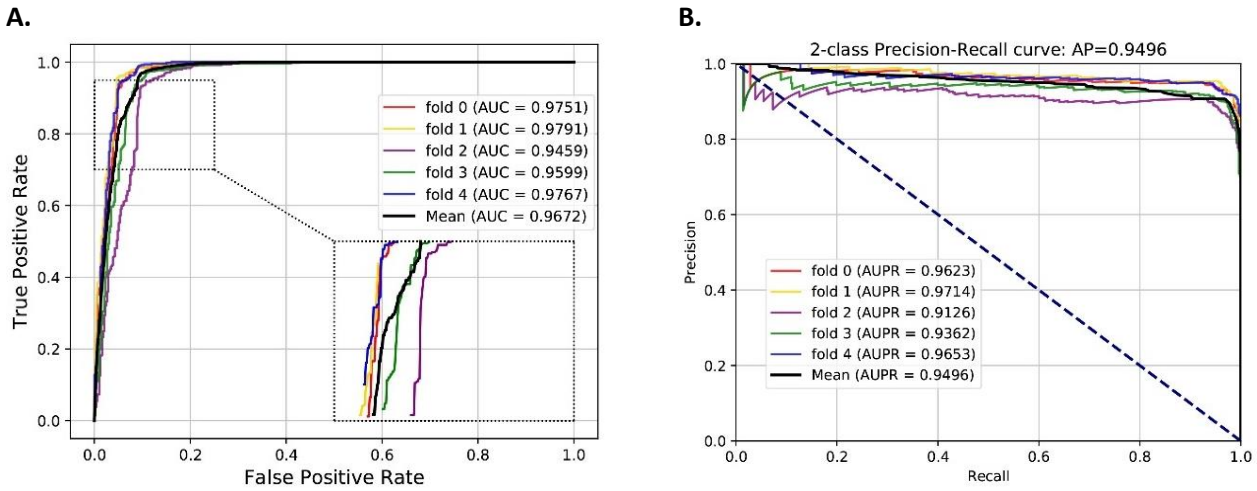

Figure 3. ROC curves (A) and PR curves (B) generated by the DWPREPHI model on the ESKAPE dataset based on 5-fold cross-validation.

Table 2. Predictive performance of the DWPREPHI model on the ESKAPE model.

Classifiers	ACC. (%)	Sen. (%)	Spec. (%)	Prec. (%)	F1	AUC
RF	84.07±1.54	83.22±2.37	84.92±0.97	84.65±1.10	83.92±1.67	0.9153±0.0133
SVM	83.46±0.88	78.84±2.29	88.08±0.80	86.88±0.56	82.64±1.16	0.9085±0.0096
KNN	72.66±0.49	82.76±2.04	62.57±2.31	68.87±0.85	75.16±0.57	0.8333±0.0067
GBDT	85.03±4.33	97.38±1.15	72.68±8.08	78.34±5.24	86.76±3.43	0.9109±0.0478
DNN	92.25±1.52	98.70±0.74	85.80±2.33	87.45±1.88	92.73±1.37	0.9672±0.0142

including random forest (RF) [44], support vector machine (SVM) [45], K Nearest Neighbors (KNN) [46], and Gradient Boosting Decision Tree (GBDT) [47]. As with the models presented in the paper, we also used a 5-fold cross-validation approach, and the specific prediction results for these four methods were presented in Table 2. The results showed that the GBDT-based method achieved the best results among the four models with an accuracy and AUC of 85.03% and 0.9109, respectively, but is still 7.22% and 0.0536 lower than our model in terms of accuracy and AUC. To provide a more intuitive comparison, we presented all the comparison results in the form of bar charts in Figure 4. The combined comparison results proved that the traditional machine learning based classifier predicted somewhat lower results than the deep learning-based classifier. This could be attributed to the fact that deep neural networks could capture the complex non-linear relationships between input and output data.

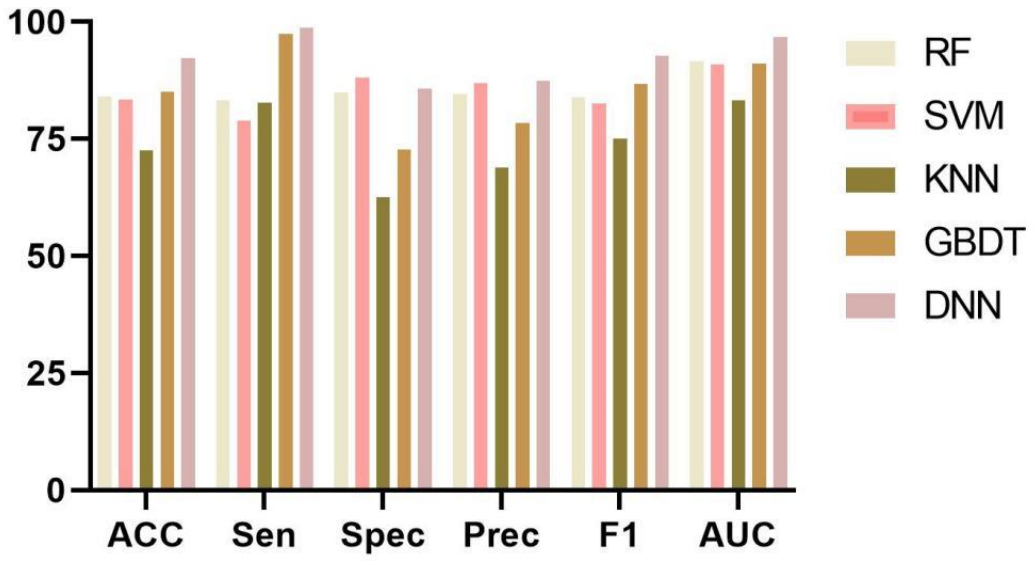

Figure 5. Comparison of the performance of different classifiers.

Table 3. Comparison of prediction performance with different network embedding algorithms.

classifiers	ACC. (%)	Sen. (%)	Spec. (%)	Prec. (%)	F1	AUC
SDNE	81.79±1.26	87.47±1.12	76.1±1.72	78.56±1.36	82.77±1.14	0.8814±0.0142
Lap	72.33±8.40	77.97±7.84	66.69±9.59	70.19±7.89	73.85±7.78	0.8071±0.0762
Hope	74.46±7.84	78.38±9.38	70.55±7.13	72.61±6.71	75.34±7.8	0.8339±0.0678
This study	**92.25±1.52**	**98.70±0.74**	**85.80±2.33**	**87.45±1.88**	**92.73±1.37**	**0.9672±0.0142**

Comparison between different network embedding algorithms

To evaluate the effectiveness of the deep wandering algorithm, we compared it with a number of popular network embedding algorithms including SDNE (Structural Deep Network Embedding) [48], Lap (Laplacian Eigenmaps) [49], and Hope. For a fair comparison, all methods were trained to predict with the same data dimension and DNN structure. Table 3 presents the prediction results of these three powerful graph embedding algorithms. The results showed that all methods perform lower than the depth wandering based methods. The prediction performance of the two factorization-based methods (Lap and Hope) was generally 5-9% higher than that of the deep learning-based algorithm (SDNE). However, despite the good prediction results of SDNE, it was still 10.46% lower than our method in terms of ACC values, which suggested that our Deepwalk algorithm based on the random walk principle could improve the prediction performance of the model in this experiment.

Case study

To further assess the realistic performance of the proposed model for predicting phage-host interactions, we did case studies for three pathogenic strains of *E. coli*, *Pseudomonas aeruginosa*, and *S. enterica*. We used the known ESKAPE dataset as a training set to make predictions for all three possible phage-host relationships. The top 10 pairs with the highest prediction scores were then selected for query validation in the EMBL-EBL database. The results were shown in Tables 4-6, where 7, 8, and 8 of the top 10 predicted phage-host interactions were experimentally validated in the experimental data provided by the EMBL-EBL database, respectively.

Table 4. The top 10 phages predicted by the proposed model to be associated with *E. coli*.

Rank	EMBL-EBL ID	Evidence	Rank	EMBL-EBL ID	Evidence
1	AJ505988	Confirmed	6	AP009390	N.A.
2	Z36986	Confirmed	7	EU330206	Confirmed
3	X06792	Confirmed	8	FJ839693	Confirmed
4	X04442	Confirmed	9	GU323318	N.A.
5	X01753	Confirmed	10	KM501444	N.A.

Table 5. The top 10 phages predicted by the proposed model to be associated with *P. aeruginosa*.

Rank	EMBL-EBL ID	Evidence	Rank	EMBL-EBL ID	Evidence
1	AM265638	Confirmed	6	LN610573	Confirmed
2	HG518155	Confirmed	7	KC862296	Confirmed
3	FN263372	Confirmed	8	KF147891	Confirmed
4	KX587949	N.A.	9	KU948710	Confirmed
5	GU988610	Confirmed	10	KT001918	N.A.

Table 6. The top 10 phages predicted by the proposed model to be associated with *S. enterica*.

Rank	EMBL-EBL ID	Evidence	Rank	EMBL-EBL ID	Evidence
1	MF001354	Confirmed	6	EF151188	Confirmed
2	EF212166	Confirmed	7	GU573886	Confirmed
3	MH709120	Confirmed	8	MF188997	Confirmed
4	MF415410	N.A.	9	CP018657	Confirmed
5	KY652726	N.A.	10	CP000026	Confirmed

Conclusion

In this study, we proposed a model for predicting potential phage-host interactions based on a graph embedding algorithm. To uncover the hidden relationships between phages and hosts, the model fully combined information on link behavior and node attributes of phage-host interactions graphs and effectively predicted the relationships between phages and hosts using deep neural networks as classifiers. Results of cross-validation on the ESKAPE dataset showed that the model had excellent overall predictive performance. The model also achieved optimal results in comparison with different machine learning-based classifiers and graph embedding algorithms. In addition, to further demonstrate the practical value of the model, a case study of three pathogenic bacteria (*E. coli*, *Pseudomonas aeruginosa*, and *S. enterica*) was conducted, and the prediction results were supported by relevant experiments and databases. Taken together, these experimental results showed that our proposed DWPREPHI model was reliable in predicting phage-host interactions and could provide potential phages for biological experiments, offering a new option for phage therapy. In future studies, we will try to incorporate phage-host similarity information and semantic information and optimize the prediction framework in anticipation of achieving better prediction results.

References

1. Angermeyer A, Hays SG, Nguyen M, Johura FT, Sultana M, Alam M, *et al*. 2022. Evolutionary sweeps of subviral parasites and their phage host bring unique parasite variants and disappearance of a phage CRISPR-Cas system. Mbio. 13(1):e03088-e4021.

2. Federici S, Kviatcovsky D, Valdés-Mas R, Elinav E. 2023. Microbiome-phage interactions in inflammatory bowel disease. Clin Microbiol Infect. 29(6):682-688.
3. Gubatan J, Boye TL, Temby M, Sojwal RS, Holman DR, Sinha SR, *et al*. 2022. Gut microbiome in inflammatory bowel disease: role in pathogenesis, dietary modulation, and colitis-associated colon cancer. Microorganisms. 10(7):1371.
4. Lai JY, Lim TS. 2020. Infectious disease antibodies for biomedical applications: A mini review of immune antibody phage library repertoire. Int J Biol Macromol. 163:640-648.
5. Duong MT, Qin Y, You SH, Min JJ. 2019. Bacteria-cancer interactions: bacteria-based cancer therapy. Exp Mol Med. 51(12):1-15.
6. Islam SU, Ul-Islam M, Ahsan H, Ahmed MB, Shehzad A, Fatima A, *et al*. 2021. Potential applications of bacterial cellulose and its composites for cancer treatment. Int J Biol Macromol. 168:301-309.
7. Sedighi M, Bialvaei AZ, Hamblin MR, Ohadi E, Asadi A, Halajzadeh M, *et al*. 2019. Therapeutic bacteria to combat cancer: current advances, challenges, and opportunities. Cancer medicine. 8(6):3167-3181.
8. Naik RK, Naik MM, D'Costa PM, Shaikh F. 2019. Microplastics in ballast water as an emerging source and vector for harmful chemicals, antibiotics, metals, bacterial pathogens and HAB species: A potential risk to the marine environment and human health. Mar Pollut Bull. 149:110525.
9. Christaki E, Marcou M, Tofarides A. 2020. Antimicrobial resistance in bacteria: mechanisms, evolution, and persistence. J Mol Evol. 88(1):26-40.
10. Frost HM, Munsiff SS, Lou Y, Jenkins TC. 2022. Simplifying outpatient antibiotic stewardship. Infect Control Hosp Epidemiol. 43(2):260-261.
11. Serwecińska L. 2020. Antimicrobials and antibiotic-resistant bacteria: a risk to the environment and to public health. Water. 12(12):3313.
12. Miethke M, Pieroni M, Weber T, Brönstrup M, Hammann P, Halby L, *et al*. 2021. Towards the sustainable discovery and development of new antibiotics. Nat Rev Chem. 5(10):726-749.
13. Nale JY, Clokie MR. 2021. Preclinical data and safety assessment of phage therapy in humans. Curr Opin Biotechnol. 68:310-317.
14. Pan J, You W, Lu X, Wang S, You Z, Sun Y. 2023. GSPHI: a novel deep learning model for predicting phage-host interactions via multiple biological information. Comput Struct Biotechnol J. 21:3404-341.
15. Leite DMC, Brochet X, Resch G, Que YA, Neves A, Peña-Reyes C. 2018. Computational prediction of inter-species relationships through omics data analysis and machine learning. BMC bioinformatics. 19:151-159.
16. Leite DMC, Lopez JF, Brochet X, Barreto-Sanz M, Que YA, Resch G, *et al*. 2018. Exploration of multiclass and one-class learning methods for prediction of phage-bacteria interaction at strain level. IEEE Inte Conf Bioinfo Biomed (BIBM). Madrid, Spain. 2018:1818-1825.
17. Zhu Q, Dai Q, He R, Huang J. 2022. PHIHNE: predicting Phage-Host Interaction through Heterogeneous Network Embedding. IEEE/WIC/ACM Inte Joint Conf Web Intel Intelligent Agent Technol (WI-IAT). Niagara Falls, ON, Canada. 2022:914-921.
18. Zhou F, Gan, R, Zhang F, Ren C, Yu L, Si Y, *et al*. 2022. PHISDetector: A tool to detect diverse in silico phage–host interaction signals for virome studies. Genomics, Proteomics Bioinf. 20(3):508-523.
19. Galiez C, Siebert M, Enault F, Vincent J, Söding J. 2017. WIsH: who is the host? Predicting prokaryotic hosts from metagenomic phage contigs. Bioinformatics. 33(19):3113-3114.
20. Pakhrin SC, Shrestha B, Adhikari B, Kc DB. 2021. Deep learning-based advances in protein structure prediction. Int J Mol Sci. 22(11):5553.
21. Villarroel J, Kleinheinz KA, Jurtz VI, Zschach H, Lund O, Nielsen M, *et al*. 2016. HostPhinder: a phage host prediction tool. Viruses. 8(5):116.
22. Rube HT, Rastogi C, Feng S, Kribelbauer JF, Li A, Becerra B, *et al*. 2022. Prediction of protein–ligand binding affinity from sequencing data with interpretable machine learning. Nat Biotechnol. 40(10):1520-1527.
23. Chen ZH, You ZH, Guo ZH, Yi HC, Luo GX, Wang YB. 2020. Prediction of drug–target interactions from multi-molecular network based on deep walk embedding model. Front Bioeng Biotechnol. 8:338.
24. Li G, Luo J, Wang D, Liang C, Xiao Q, Ding P, *et al*. 2020. Potential circRNA-disease association prediction using DeepWalk and network consistency projection. J biomed informat. 112:103624.
25. Wong L, You ZH, Guo ZH, Yi HC, Chen ZH, Cao MY. 2020. MIPDH: a novel computational model for predicting microRNA–mRNA interactions by DeepWalk on a heterogeneous network. ACS omega. 5(28):17022-17032.
26. Li M, Zhang W. 2022. PHIAF: prediction of phage-host interactions with GAN-based data augmentation and sequence-based feature fusion. Brief Bioinformat. 23(1):bbab348.
27. Ren X, Li XY, Ren KJ, Song JQ, Xu ZC, Deng KF, *et al*. 2021. Deep learning-based weather prediction: a survey. Big Data Res. 23:100178.
28. Putka DJ, Beatty AS, Reeder MC. 2018. Modern prediction methods: New perspectives on a common problem. Organizat Res Meth. 21(3):689-732.
29. Boutet E, Lieberherr D, Tognolli M, Schneider M, Bansal P, Bridge AJ, *et al*. 2016. UniProtKB/Swiss-Prot, the manually annotated section of the UniProt KnowledgeBase: how to use the entry view. Plant bioinformatics: methods and protocols. 1374:23-54.
30. Suzek BE, Huang H, McGarvey P, Mazumder R, Wu CH. 2007. UniRef: comprehensive and non-redundant UniProt reference clusters. Bioinformatics. 23(10):1282-1288.
31. Ayobami O, Brinkwirth S, Eckmanns T, Markwart R. 2022. Antibiotic resistance in hospital-acquired ESKAPE-E infections in low-and lower-middle-income countries: a systematic review and meta-analysis. Emerg Microbes Infect. 11(1):443-451.
32. Boeckaerts D, Stock M, Criel B, Gerstmans H, Baets BD, Briers Y. 2021. Predicting bacteriophage hosts based on sequences of annotated receptor-binding proteins. Sci Rep. 11(1):1467.
33. Perozzi B, Al-Rfou R, Skiena S. 2014. Deepwalk: Online learning of social representations. Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining. Association for Computing Machinery, New York, NY, USA. 2014:701-710.
34. Ma L, Zhang Y. 2015. Using Word2Vec to process big text data. IEEE International Conference on Big Data (Big Data), Santa Clara, CA, USA. 2015: 2895-2897.
35. Onishi T, Shiina H. 2020. Distributed representation computation using CBOW model and skip–gram model. 9th International Congress on Advanced Applied Informatics (IIAI-AAI). Kitakyushu, Japan. 2020:845-846.
36. Tsukiyama S, Hasan MM, Fujii S, Kurata H. 2021. LSTM-PHV: prediction of human-virus protein–protein interactions by LSTM with word2vec. Brief Bioinformat. 22(6):bbab228.
37. Sun W, Su F, Wang L. 2018. Improving deep neural networks with multi-layer maxout networks and a novel initialization method. Neurocomputing. 278:34-40.
38. Agarap AF. 2018. Deep learning using rectified linear units (relu). arXiv:1803.08375.

39. Li Z, Li H, Jiang X, Chen B, Zhang Y, Du G. 2018. Efficient FPGA implementation of softmax function for DNN applications. 12th IEEE International Conference on Anti-counterfeiting, Security, and Identification (ASID). Xiamen, Fujian, China. 2018:212-216.
40. Ruby U, Yendapalli V. 2020. Binary cross entropy with deep learning technique for image classification. Int J Adv Trends Comput Sci Eng. 9(4):5393-5397.
41. Baldi P, Sadowski P. 2014. The dropout learning algorithm. Artif Intell. 210:78-122.
42. Kingma DP, Ba J. 2014. Adam: A method for stochastic optimization. arXiv:1412.6980.
43. Fushiki T. 2011. Estimation of prediction error by using K-fold cross-validation. Stat Comput. 21:137-146.
44. Rigatti S. 2017. Random forest. J Insur Med. 47(1):31-39.
45. Noble W. 2006. What is a support vector machine? Nat biotech. 24(12):1565-1567.
46. Kramer O. 2013. Dimensionality reduction with unsupervised nearest neighbors. Springer Berlin, Heidelberg, Germany. pp. 13-23.
47. Ke G, Meng Qi, Finley T, Wang T, Chen Wei, Ma W, *et al*. 2017. Lightgbm: A highly efficient gradient boosting decision tree. Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17). Curran Associates Inc., Red Hook, NY, USA. 2017:3149–3157.
48. Wang D, Cui P, Zhu W. 2016. Structural deep network embedding. Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining. Association for Computing Machinery, New York, NY, USA. 2016:1225-1234.
49. Chen Z, Chen C, Zhang Z, Zheng Z, Zou Q. 2019. Variational graph embedding and clustering with Laplacian Eigenmaps. Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI'19). AAAI Press. 2019:2144–2150.