OXFORD

*Briefings in Bioinformatics*, 2025, **26(1)**, bbae672

https://doi.org/10.1093/bib/bbae672

**Problem Solving Protocol**

# Predicting phage–host interactions via feature augmentation and regional graph convolution

Ankang Wei[1,2,3], Zhen Xiao[1,2,3], Lingling Fu[1,2,3], Weizhong Zhao[2,3,4], Xingpeng Jiang[2,3,4,*]

[1]School of Mathematics and Statistics, Central China Normal University, Wuhan 430079, China
[2]Hubei Provincial Key Laboratory of Artificial Intelligence and Smart Learning, Central China Normal University, Wuhan 430079, China
[3]School of Computer Science, Central China Normal University, Wuhan 430079, China
[4]National Language Resources Monitoring & Research Center for Network Media,Central China Normal University, Wuhan 430079, China

[*]Corresponding author. E-mail: xpjiang@mail.ccnu.edu.cn

## Abstract

Identifying phage–host interactions (PHIs) is a crucial step in developing phage therapy, which is the promising solution to addressing the issue of antibiotic resistance in superbugs. However, the lifestyle of phages, which strongly depends on their host for life activities, limits their cultivability, making the study of predicting PHIs time-consuming and labor-intensive for traditional wet lab experiments. Although many deep learning (DL) approaches have been applied to PHIs prediction, most DL methods are predominantly based on sequence information, failing to comprehensively model the intricate relationships within PHIs. Moreover, most existing approaches are limited for sub-optimal performance, due to the potential risk of overfitting induced by the highly data sparsity in the task of PHIs prediction. In this study, we propose a novel approach called MI-RGC, which introduces mutual information for feature augmentation and employs regional graph convolution to learn meaningful representations. Specifically, MI-RGC treats the presence status of phages in environmental samples as random variables, and derives the mutual information between these random variables as the dependency relationships among phages. Consequently, a mutual information-based heterogeneous network is construted as feature augmentation for sequence information of phages, which is utilized for building a sequence information-based heterogeneous network. By considering the different contributions of neighboring nodes at varying distances, a regional graph convolutional model is designed, in which the neighboring nodes are segmented into different regions and a regional-level attention mechanism is employed to derive node embeddings. Finally, the embeddings learned from these two networks are aggregated through an attention mechanism, on which the prediction of PHIs is condcuted accordingly. Experimental results on three benchmark datasets demonstrate that MI-RGC derives superior performance over other methods on the task of PHIs prediction.

**Keywords**: mutual information; metagenomic data; phage–host interaction; regional graph convolutional network; regional-level attention

## Introduction

Since the discovery of penicillin by Alexander Fleming in 1928, it has been widely used to treat various bacterial infections [1]. However, the overuse of antibiotics has led to the emergence of resistant bacteria [2]. Today, antibiotic resistance in various bacteria poses a global threat to the treatment of bacterial infections [3]. The specific ability of phages to recognize and kill bacterial hosts offers a promising solution for treating and controlling antibiotic-resistant bacteria. In 2016, phages were successfully used for the first time to treat a patient infected with a multidrug-resistant strain of Acinetobacter baumannii, leading to global recognition of phage therapy [4]. As research on phages deepens, it has been discovered that phages influence the stability of microbial communities through their interactions with bacterial hosts. Phages play significant roles in the pathogenesis of diseases such as parkinson's disease [5], diabetes [6], and other diseases. Understanding and elucidating phage–host interactions (PHIs) is crucial for in-depth research on phages. However, the dependence of phages on bacterial hosts for their life cycle makes studying PHIs through traditional experimental methods extremely challenging due to stringent culture conditions, significantly limiting the progress of experimental approaches [7]. Additionally, only 1% of microbial cells in natural environments are culturable, making the available bacterial hosts for culture very limited [8].

With the development of metagenomic technology, a large amount of genetic data from phages and their host bacteria, as well as metagenomic data, have been clarified. Predicting the interactions between phages and bacteria using computational methods is highly significant [9]. During the co-evolution of phages and hosts, gene exchange may occur, or molecular signals such as Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR) may be generated during the infection and resistance process. These pieces of information can be used to compute PHIs, thereby greatly reducing the time and cost required to elucidate PHIs [10]. Using computational models not only help to clarify PHIs but also predict the host range of newly discovered phages and provide reliable phages for new bacteria [11].

**Received:** August 5, 2024. **Revised:** November 5, 2024. **Accepted:** December 14, 2024
© The Author(s) 2024. Published by Oxford University Press.
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (https://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

The predictive performance of current computational models primarily depends on feature extraction and model selection. From the perspective of feature extraction, existing models can be broadly classified into those based on DNA sequence homology, nucleotide or amino acid sequence features, and protein-based features along with their physicochemical properties [12]. Exploiting features that reflect the relationship between phages and hosts is crucial for PHIs prediction. For instance, prediction methods such as VHM [13], PHP [14], and DeepHost [15] utilize k-mer frequency from DNA sequences; HostG [16] and CHERRY [17] employ CRISPR sequences; PHIAF [18] extracts features from both DNA sequences and various physicochemical properties of proteins. Additionally, features like receptor-binding protein (RBP) used by Boeckaerts [19], and HMM profiles used by RaFAH [20] and vHULK [21], are also noteworthy examples.

From the perspective of model design, existing prediction methods can be mainly divided into statistical models, machine learning models, deep learning (DL) models, graph models, and hybrid models [12]. Statistical models primarily predict hosts by calculating the match or dissimilarity between phage and bacterial gene sequences. For instance, VHM [13] is one of the earliest methods that used k-mer frequency dissimilarity for phage host range prediction, and WisH [22] utilizes the Markov model. Machine learning models can be categorized into supervised and unsupervised learning models. For example, the PHIDetector [23] model trains decision trees, logistic regression, support vector machines, Gaussian Naive Bayes, and Bernoulli Naive Bayes models. Other machine learning models include random forests, and gradient boosting [12]. Deep learning models simplify the feature extraction and model design processes, making them suitable for large-scale data and high-dimensional features. Examples include the Convolutional Neural Network (CNN) used by ContigNet [24] and the Generative Adversarial Network used by PHIAF [18]. Graph models can simulate the potential interactions between phages and their bacterial hosts, offering greater adaptability. HostG [16] and CHERRY [17] are examples of prediction methods that construct networks, a strategy currently yielding the best results. Hybrid models combine statistical models with machine learning models, such as RaFAH [20], or statistical models with DL models, such as HoPhage [25].

In the task of PHIs prediction, although several predictive methods have achieved promising results, there are still some challenges. Among the currently available types of information, most, aside from sequence information, have very limited data volume and are applicable only to specific datasets, such as CRISPR [26], RBP [27], Receptor Protein [28], and Quorum Sensing [29]. It is essential to explore more informative and broadly applicable data. Additionally, existing methods that use graph neural networks can lead to information loss and redundancy during data integration, as they do not account for the varying contributions of neighbors at different distances to the aggregation of central node information, thus diminishing model performance.

In response to these challenges, we propose a mutual information-based augmentation module and a Regional Graph Convolutional (RGC) module. First, considering the survival characteristics of phages that depend on bacterial hosts for their life activities, we aim to extract more reliable information from the microbial environment. We utilize mutual information from information theory to explore and evaluate the correlations of phages in metagenomic data. By treating the expression information of phages in various environmental samples as random variables, we quantify the information shared between any two variables, thereby constructing a mutual information network for phages to enhance their features.

Secondly, to better account for the contributions of neighboring nodes at different distances to the central node, we introduce the RGC model. In many prediction tasks based on graph models, researchers construct heterogeneous networks using various distance metrics, but often neglect the varying contributions of nodes at different distances when aggregating information. This oversight inevitably generates a large amount of redundant information. In such cases, directly applying graph convolution, graph attention, or other graph-based models for feature learning is inadvisable. Some existing methods enhance features by setting a threshold to eliminate low-similarity connections. While this operation improves the predictive performance of the model, it also introduces a risk of information loss to some extent.

In this study, we propose a novel end-to-end prediction model (MI-RGC) for PHI prediction and conduct experiments on three datasets, with results outperforming the latest baseline models. The contributions of this study are as follows:

- We explored dependencies among phages in metagenomic data for the first time using the concept of mutual information.
- Sequence features were enhanced by integrating mutual information among phages.
- Developed a regional graph convolutional model that learns from densely connected sequence-based heterogeneous graphs while using regional-level attention to learn the contribution of neighbors from different regions to the central node.
- This paper proposes a novel end-to-end prediction model and has conducted experiments on five datasets.

## Materials and methods

### Dataset

In this study, we used one metagenomic dataset and three PHI-related datasets for experiments. The metagenomic data used for mutual information calculations contain 370 samples from the human gut environment and is available for download under PRJNA422434 [30].

The three PHI datasets are the ours-dataset, CHERRY-dataset [17], and PHD-dataset [31]. The ours-dataset is a custom dataset we constructed, containing entities with explicit abundance and sequence information. The CHERRY-dataset is one of the most widely used datasets in existing methods; we used it to perform comparative experiments to better assess the predictive performance of our model. Additionally, to evaluate the stability of model performance, we constructed a larger dataset compared to the ours-dataset and CHERRY-dataset for comparative analysis. In the experiments, all three datasets were structured as bipartite graphs, which primarily include two types of nodes, phages and hosts, and edge relationships constructed based on PHIs. The classification information of the bacteriophages is referenced from the International Committee on Taxonomy of Viruses. Table 1 presents the number of PHIs across three datasets, including the number of phages at the species level (Phage (S)), the number of phages at the genus level (Phage (G)), the number of hosts at the species level (Host (S)), and the number of hosts at the genus level (Host (G)).

Table 1. Summary statistics for all datasets

| datasets | PHIs | Phage(S) | Phage(G) | Host(S) | Host(G) |
|---|---|---|---|---|---|
| Our | 2002 | 1909 | 414 | 278 | 128 |
| CHERRY | 1876 | 1876 | 109 | 207 | 111 |
| PHD | 4724 | 4724 | 1594 | 553 | 149 |

## Data preparation

This section mainly introduces how we processed sequence information and obtained the mutual information of phages.

### *Computation of sequence information*

We downloaded the gene sequences of phages and computed their *K*-mer features. Then, by calculating the cosine similarity of *K*-mer ($k = 3$) features between phages, we obtained a similarity network among phages.

$$S_{ij} = \frac{\overrightarrow{p_i} \cdot \overrightarrow{p_j}}{\|p_i\|\|p_j\|} \tag{1}$$

where $\overrightarrow{p_i}$ and $\overrightarrow{p_j}$ represent the characteristics of phage $p_i$ and $p_j$, respectively. $\|\cdot\|$ represents the $\ell_2$ norm of the vector.

We assumed the relationship network between hosts to be discrete, setting the adjacency matrix between hosts to be an identity matrix in network construction. This is because the proportion of phage-related information in host sequences is very small, which is not conducive to prediction. Based on the bipartite network constructed according to associations, we obtained a heterogeneous network based on sequence information.

### *Computation of mutual information*

Mutual information, an important concept in information theory, is used to quantify the correlation and dependency between two random variables [32]. Based on the characteristic that phages rely on host bacteria for their life activities, we hypothesize that if two phages share a significant amount of information across multiple environments, they are likely to have similar host preferences. Under this basic assumption, we used metagenomic analysis tools to calculate the presence status of phages in various samples.

First, we utilized KneadData [33] to filter out low-quality reads and adapter sequences from the metagenomic data. Subsequently, FastQC [34] was employed to assess the quality of the filtered data. We then used Kraken2 [35] for taxonomic classification and annotation of the DNA sequences to identify microbial sequences, with Bracken [36] used to enhance the accuracy of species-level annotations. Finally, we obtained the absolute abundance of phages and bacteria across 370 samples, determining the presence and distribution of phages and bacteria within each sample.

The presence of phages in various samples is dependent on the presence of their host bacteria. When two phages exhibit a high degree of correlation or dependency in their state information, the primary cause of this high correlation is likely the significant similarity or even identity of their corresponding hosts. We consider the presence state of phages across different samples as random variables, and calculating the association between phages involves computing the mutual information of these corresponding random variables. Below are some formulas related to the calculation of mutual information [37].

Assume there is a discrete variable $X$ with its probability distribution denoted as $p(X)$. The information entropy $H(X)$ is defined as

$$H(X) = -\sum_{x \in X} p(x) log p(x) \tag{2}$$

For random variables $X$ and $Y$, their entropy is denoted as $H(X)$ and $H(Y)$ respectively. The conditional entropy $H(Y|X)$ are defined as follows:

$$H(Y|X) = \sum_{x \in X} p(x) H(Y|X = x) \tag{3}$$

$$= -\sum_{x \in X} p(x) \sum_{y \in Y} p(y|x) log p(y|x) \tag{4}$$

$$= -\sum_{x \in X} \sum_{y \in Y} p(y, x) log p(y|x) \tag{5}$$

where $p(y, x) = p(x, y)$ represents the probability of events $Y = y$ and $X = x$ occurring simultaneously. The following is the formula for calculating mutual information $I(X; Y)$:

$$I(X; Y) = H(Y) - H(Y|X) \tag{6}$$

$$= H(X) - H(X|Y) \tag{7}$$

To better assess the degree of association between random variables, we have normalized mutual information using a method called Normalized Mutual Information (NMI). The normalized result falls within the range of $[0, 1]$. When the result approaches 1, it indicates a stronger association between the variables, and when it approaches 0, it indicates a weaker association, implying that the variables are independent of each other.

$$NMI = \frac{I(X; Y)}{\sqrt{H(X) \cdot H(Y)}} \tag{8}$$

We have obtained a phage association network based on mutual information. This network determines whether there is an association between phages, and the strength of the association is determined by their corresponding similarity measures. Based on the bipartite network constructed according to associations, we obtained a heterogeneous network based on mutual information.

We constructed two types of homogenous networks for phages, one based on phage sequence information and the other on mutual information between phages. Combined with the bipartite network constructed from association data, we ultimately generated two types of heterogeneous networks representing PHIs.

## Regional graph convolution

Considering the varying contributions of neighbors at different distances to the central node, this study segments the neighbors of the central node into distinct intervals based on distance and introduces a Regional-Level Attention mechanism (RL-AT)

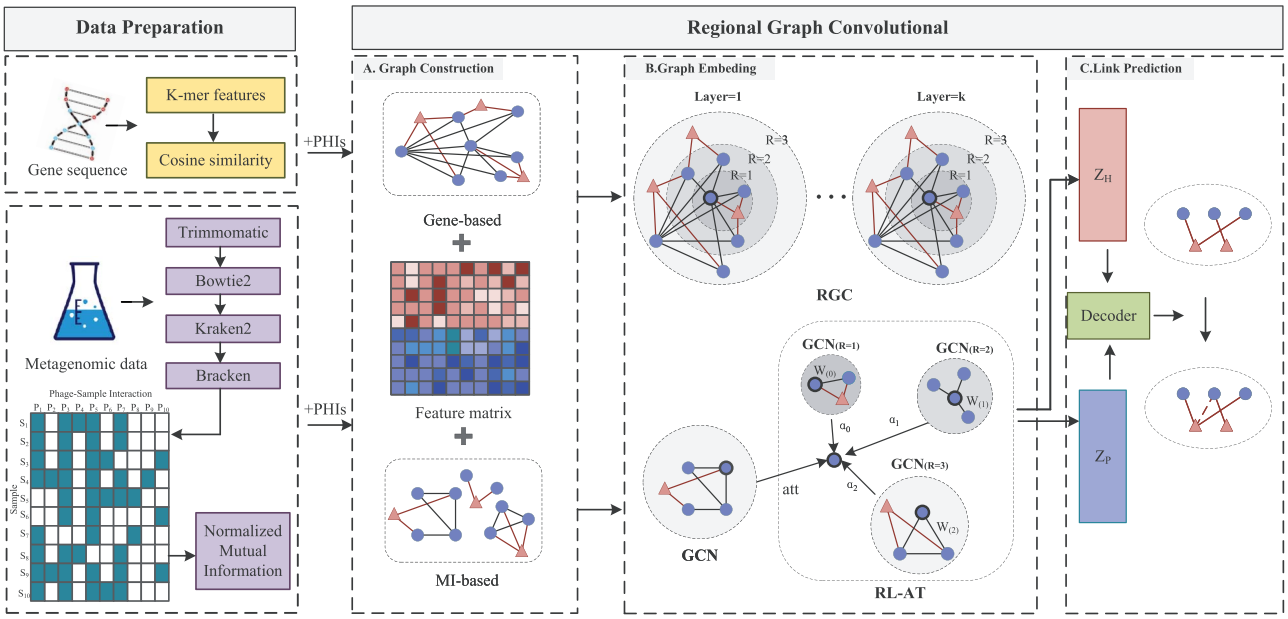



Figure 1. The overall framework of MI-RGC consists of two main components: data preparation and Regional Graph Convolution. The data preparation section is further divided into two subsections: Dataset and Data Preparation. The Regional Graph Convolutional section is divided into three subsections: (A) Graph Construction; (B) Graph Embedding; and (C) Link Prediction.

to derive more meaningful node embeddings. Subsequently, an inner product decoder is employed on the resultant embeddings to predict PHIs. Our model is trained in an end-to-end manner, with parameters updated through gradient descent to minimize the loss function. The overall workflow of the model is illustrated in Fig. 1.

## Graph construction

We denote the two heterogeneous networks constructed based on sequence information and mutual information as $G_{sq}$ and $G_{mi}$, respectively. To facilitate feature integration in the later stages, we set the initial node representations of both networks to be the same.

$$G_{sq} = \begin{pmatrix} P_m & A_{m\times N} \\ A^T_{m\times N} & H_n \end{pmatrix} \tag{9}$$

$P_m$ denotes the adjacency matrix of phages based on similarity measures, with $m$ representing the number of phages. $H_n$ is the identity matrix, with $n$ representing the number of hosts. $A_{m\times N}$ stands for the association matrix between phages and hosts, derived from known associations. For a phage $i$ and host $j$, $Aij = 1$ if they are associated; otherwise, $Aij = 0$.

To prevent redundant learning of topological information, we set the initial node features as follows:

$$\begin{pmatrix} 0 & trainA \\ trainA^T & 0 \end{pmatrix} \tag{10}$$

$trainA$ represents the association matrix of the training set.

$$G_{mi} = \begin{pmatrix} P'_m & A_{m\times N} \\ A^T_{m\times N} & H_n \end{pmatrix} \tag{11}$$

here, $P'_m$ denotes the adjacency matrix of phages, and its values are determined by both mutual information and similarity measures.

## Graph embeding

In the graph convolution model, the local computation graph is a fundamental concept in graph convolutional networks, focusing on the central node and its neighbors for feature aggregation and information propagation. Parameters and weights are shared across all local computation graphs, and the same information propagation method should be used within the same local computation graph. As illustrated in Fig. 1, we categorize the neighbors of each node in the graph into different regions based on distance metrics. Each region, along with the central node, forms a subgraph that can be regarded as a local computation graph. The figure demonstrates three local computation graphs, and in our actual experiments, we set the number of local computation graphs to four. Next, we will present the processes for regional representation of the graph and node embedding calculations.

For the graph G, we label the neighbors of the central node based on distance metrics among the nodes in the network, dividing all neighbors into different regions to obtain a multi-region network $G = \{V, E, \mathcal{R}, \varepsilon\}$, where V represents a set of nodes, $E$ represents a set of edges, $\mathcal{R}$ represents a set of regions partitioned based on similarity, and $\varepsilon$ represents the interval length for region partitioning, where $\mathcal{R} = \{r_1, r_2, ..., r_m\}$, where each region $r_i$ represents the set of neighbor nodes whose distances from the central node fall within the interval $[1-(i-1)*\varepsilon, i*\varepsilon]$. The nodes in each region, along with the central node, collectively form a local computation graph. The number of regions corresponds to the number of local computation graphs established within this framework.The embedding process of nodes during the information aggregation and propagation in a single-layer GCN can be represented as follows:

$$Z_i = f(Z_{i,0}, Z_{i,1}, ..., Z_{i,t}) \tag{12}$$

here, $Z_{i,r}$, $0 \le r \le t$, represents the hidden representation of the node $i$ within the computational graph of the $r$-th region, and $f(\cdot)$ denotes the aggregation function.

The core of the information aggregation process still utilizes the graph convolution algorithm. The information aggregation rule for GCN is as follows:

$$Z_i^{(l+1)} = \sigma\left(\sum_{j\in\mathcal{N}^i} W^{(l)} Z_j^{(l)}\right) \tag{13}$$

where $Z_i^{(l)}$ represents the hidden layer embedding of node $i$ at layer $l$ and $W^{(l)}$ denotes the parameter matrix. $\mathcal{N}^i$ represents the neighboring nodes and $\sigma$ represents the activation function ReLU.

Each layer of the GCN will propagate and aggregate information based on the pre-partitioned regions. The aggregation method for a single-region GCN is as follows:

$$Z_i^{(l+1)} = \sigma\left(\sum_r \sum_{j\in\mathcal{N}_r^i} W_r^{(l)} Z_{j,r}^{(l)}\right) \tag{14}$$

where $r$ represents the partitioned region, $0 \le r \le \mathcal{R}$, and $W_r^{(l)}$ denotes the parameter matrix. The weights are shared when aggregating nodes within the same region. $\mathcal{N}_r^i$ represents the neighboring nodes within region $r$.

Region-level attention learns different weights for each region's contribution to the central node. Firstly, a linear transformation is applied to the generated representation of each region. Then, a *tanh* activation function is used to learn nonlinear features. Subsequently, another linear transformation is performed, followed by a *softmax* operation to obtain the final scores.

$$\alpha_{i,r} = \delta[W_{r_1}^{(l)} \sigma(W_{r_2}^{(l)} Z_{i,r}^{(l)})] \tag{15}$$

here, $\delta$ and $\sigma$ represent the activation functions *softmax* and *tanh*, respectively. $Z_{i,r}^{(l)}$ represents the hidden representation for the node $i$ in region $r$ of the $l$-*th* hidden layer, and $W_{r_1}^{(l)}$ and $W_{r_2}^{(l)}$ are the parameter matrices corresponding to the two linear transformations.

According to the design of region-level attention, information aggregation is performed on various regions generated around each central node. As shown in Equation 16.

$$Z_i^{(l+1)} = \sigma\left(\sum_r \sum_{j\in\mathcal{N}_r^i} \alpha_{ir} W_r^{(l)} Z_{j,r}^{(l)}\right) \tag{16}$$

Considering that the existence of edge relationships between phages in the MI-based heterogeneous graph depends on mutual information calculations, we analyzed the quantity of mutual information and found that the average degree of a viral node does not exceed 8. This indicates that the network itself is sparse, and further regional partitioning operations are unnecessary, as they would only increase the risk of model overfitting. We use the RGC model to perform feature embedding for the heterogeneous network constructed from sequence information, while utilizing $R = 0$ (*i.e.GCN*) to embed features for the heterogeneous network built on mutual information. Finally, we merge the two representations, $Z_{sq}$ and $Z_{mi}$, resulting in the final embedding, as shown in Equation 17.

$$Z = att(Z_{sq}, Z_{mi}) \tag{17}$$

### *Line-prediction*

In this study, we use an inner product decoder to predict the interaction between phages and hosts, while employing the cross-entropy loss function for model training.

$$A' = sigmoid(Z_P W' Z_H^T) \tag{18}$$

Here, $W'$ represents the trainable parameter matrix, $A'$ represents the reconstructed association matrix, and $Z_P$ and $Z_H$ represent the embedded representations of the phage and host, respectively.

The cross-entropy loss function employed in this study is especially well-suited for classification tasks.

$$L = -\frac{1}{N_P \times N_H}\left(\gamma \times \sum_{(i,j)\in y^+} log\hat{y}^{(ij)} + \sum_{(i,j)\in y^-} log(1-\hat{y}^{(ij)})\right) \tag{19}$$

Here, $y^+$ and $y^-$ represent the positive and negative sample sets, respectively. $\gamma$ is a hyperparameter introduced to mitigate the impact of the imbalance between positive and negative samples on the experiment. $N_P$ and $N_H$ denote the total numbers of phages and hosts, respectively.

Table 2. The parameters of MI-RGC

| Stryucture | Parameters |
|---|---|
| Optimizer | Adam |
| Epoch | 4000 |
| Layer | Amount(L):3 |
| dimension | Units(K):64 |
| Activation function | ReLu |
| Feature-dropout | 0.3 |
| Edge-dropout | 0.5 |
| Initial learning rate | 0.01 |
| Loss function | Cross-entropy loss function |

## Experiments

### Model basic parameter and model validation

We first conducted experiments on our own dataset to intuitively demonstrate the impact of various basic parameters on the experiments through grid tuning. These basic parameters mainly include the number of layers $L$ of graph convolution, $L$ : $\{1, 2, 3, 4, 5, 6\}$; the embedding dimension $k$ of the hidden layer, $k$ : $\{32, 64, 128, 256, 512, 1, 024\}$; the initial learning rate $lr$ of the optimizer, $lr$ : $\{0.1, 0.01, 0.001, 0.0001\}$; the number of training epochs, *epoch* : $\{500, 1000, 2000, 3000, 4000, 5000\}$; and the selection of feature-dropout and edge-dropout, *dropout* : $\{0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8\}$. Table 2 shows the best setting of the basic parameters, which is obtained by a hyper-parameter analysis.

In existing research methods for the PHIs prediction task, we commonly use area under the receiver operating characteristic curve (AUC), accuracy (ACC), area under the precision-recall curve (AUPR) and Matthews correlation coefficient (MCC) as the evaluation criteria to demonstrate experimental results. At the same time, we visualized the precision-recall (P-R) curves and receiver operating characteristic (ROC) curves for all models across the three datasets.

## Results and analysis

In this section, we conducted comprehensive experiments to compare our model with baseline models, evaluate mutual

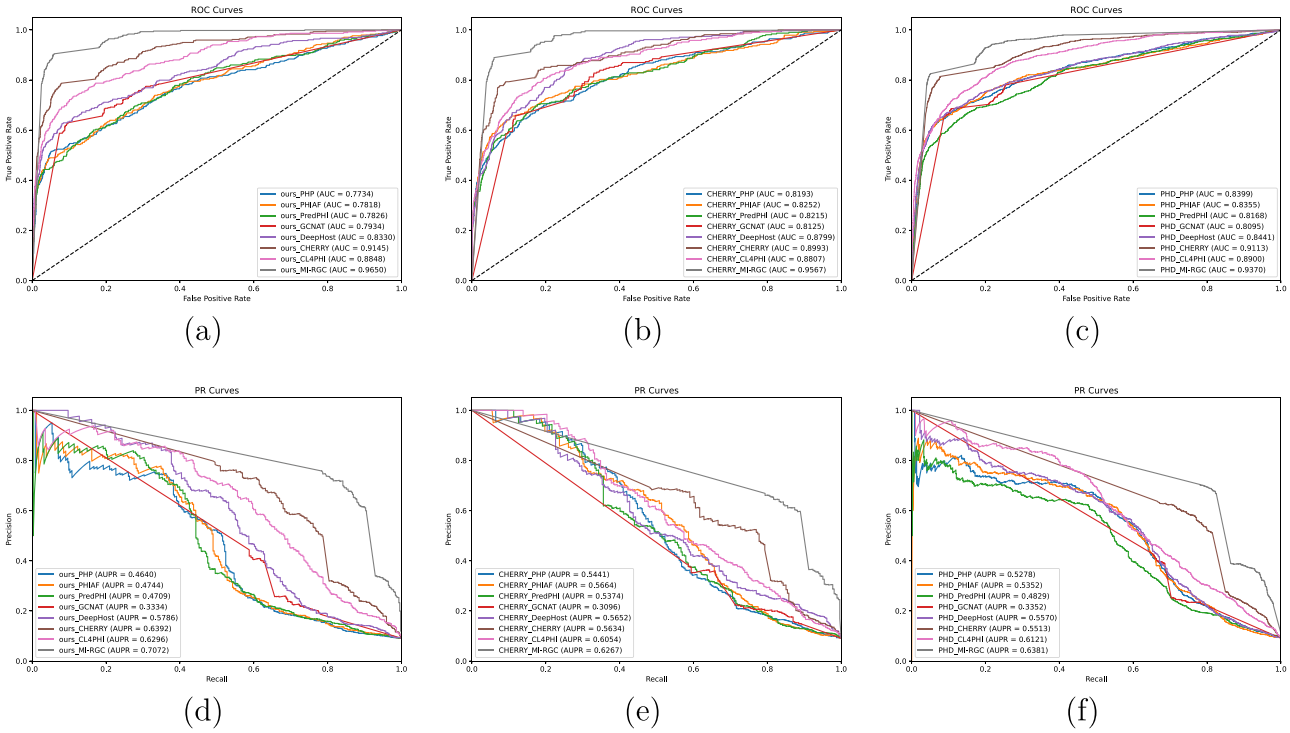



Figure 2. The visualization of the ROC curve and P-R curve for all baseline models and MI-RGC across the three datasets.

information and regional graph convolution (RGC), assess region partitioning, examine applicability across multiple datasets, and analyze parameter tuning. It is important to note that our model with the optimal settings listed in Table 2 was used in all experiments, except for those involving parameter tuning. In the comparative experiments with baseline models, all models used downsampling to balance positive and negative samples, and five-fold cross-validation was conducted on the sampled datasets. This setup ensured that all evaluation metrics were computed under conditions where the sample size and the ratio of positive to negative samples were completely consistent across the datasets.

## Comparison with baseline models

We selected the latest baseline models from recent years for comparative experiments. These models are as follows:

PredPHI [38] utilizes DL for feature embedding and employs K-means clustering to select reliable negative samples.

PHIAF [18] is the first to apply a generative adversarial network for data enhancement in PHI (PHIs) prediction tasks.

PHP [14] utilizes a Gaussian model to predict hosts for prokaryotic viruses based on k-mer frequency differences between viral and host genomes.

DeepHost [15] introduces a genome encoding method and combines it with a CNN model to predict PHIs.

CHERRY [17] constructs a knowledge graph using various protein and gene sequence features and performs PHI prediction through link prediction.

GCNAT [39] leverages graph convolutional networks and attention mechanisms to predict interactions between drugs and metabolites.

CL4PHI [40] is the first to introduce contrastive learning to improve prediction accuracy.

We conducted experiments with these baseline models and our model on the three datasets: ours-data, CHERRY-data, and PHD-data, with the results shown in Figure 2 and Table 3. As illustrated, our model outperforms the other models in all metrics. Both our model and GCNAT, as well as CHERRY, employ GCN for feature embedding. Compared to GCNAT, CHERRY's advantage lies in its use of a network built with multiple information sources, containing richer, more accurate, and higher-confidence edge relationships. The advantage of MI-RGC, beyond the feature augmentation module utilizing mutual information, is that our method effectively learns the contribution of neighbor information while also eliminating interference caused by redundant information. In Table 3, it is shown that among the baseline models, CL4PHI achieves higher metrics than PHIAF, DeepHost, PHP, and PredPHI, as CL4PHI incorporates a contrastive learning module.

## Top-K prediction accuracy

In practical applications, we aim for the predictive model to provide precise host associations for novel viruses. Therefore, we analyzed the Top-K prediction accuracy of GCNAT, CHERRY, CL4PHI, and our model. We assume that for any bacteriophage, if a validated host appears within the top-K predictions, the model's prediction for that bacteriophage is considered successful; otherwise, it is deemed a failure. The proportion of successful predictions among the total samples represents the Top-K accuracy. We selected GCNAT, CHERRY, CL4PHI, and our model for this experiment because these models are designed to support top-K output, and they incorporate distance metrics in their design. Notably, GCNAT and CHERRY utilize graph models in their framework. This metric offers a more comprehensive evaluation of the predictive performance of the models.

For this experiment, we randomly selected 50 bacteriophages from the dataset. We removed the known associations of these 50 bacteriophages from the dataset's corresponding association matrix before training the models. After training, we obtained the prediction results for these 50 bacteriophages from the result files,

Table 3. The experimental results of baseline models and MI-RGC on three datasets

| Dataset | MODEL | AUPR | AUC | ACC | MCC |
|---|---|---|---|---|---|
| ours | PHP | 0.4640 | 0.7734 | 0.9189 | 0.3217 |
| | PHIAF | 0.4744 | 0.7818 | 0.9288 | 0.4612 |
| | PredPHI | 0.4709 | 0.7826 | 0.9292 | 0.4751 |
| | GCNAT | 0.3334 | 0.7934 | 0.8466 | 0.3881 |
| | DeepHost | 0.5786 | 0.8330 | 0.9345 | 0.5122 |
| | CHERRY | 0.6392 | 0.9145 | 0.8264 | 0.4328 |
| | CL4PHI | 0.6296 | 0.8848 | 0.9257 | 0.4785 |
| | MI-RGC | 0.7072 | 0.9650 | 9300 | 0.4986 |
| CHERRY | PHP | 0.5441 | 0.8193 | 0.9307 | 0.4804 |
| | PHIAF | 0.5664 | 0.8252 | 0.9301 | 0.4663 |
| | PredPHI | 0.5374 | 0.8125 | 0.9309 | 0.4767 |
| | GCNAT | 0.3096 | 0.8799 | 0.8002 | 0.3961 |
| | DeepHost | 0.5652 | 0.8799 | 0.9258 | 0.4153 |
| | CHERRY | 0.5634 | 0.8993 | 0.8328 | 0.4435 |
| | CL4PHI | 0.6054 | 0.8807 | 0.9311 | 0.4733 |
| | MI-RGC | 0.6267 | 0.9567 | 0.9273 | 0.5163 |
| PHD | PHP | 0.5278 | 0.8399 | 0.9293 | 0.5345 |
| | PHIAF | 0.5352 | 0.8355 | 0.9299 | 0.5149 |
| | PredPHI | 0.4829 | 0.8168 | 0.9189 | 0.5001 |
| | GCNAT | 0.3352 | 0.8095 | 0.8331 | 0.3764 |
| | DeepHost | 0.5570 | 0.8441 | 0.9285 | 0.5133 |
| | CHERRY | 0.5513 | 0.9113 | 0.8870 | 0.4051 |
| | CL4PHI | 0.6121 | 0.8900 | 0.9349 | 0.5353 |
| | MI-RGC | 0.6381 | 0.9370 | 0.9240 | 0.5262 |

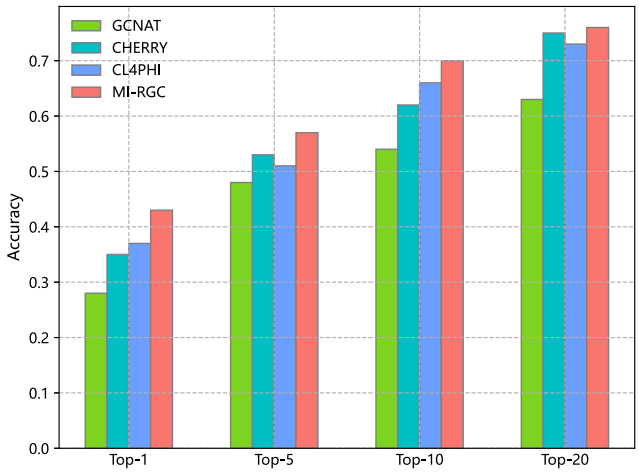



Figure 3. The experimental results of top-k prediction.

identified the Top-K predicted hosts, and calculated the accuracy accordingly. The results are shown in Fig. 3.

## Evaluation of mutual information and RGC

In this section, we designed experiments targeting mutual information and regional graph convolution. We conducted experiments by removing mutual information and replacing regional graph convolution with original graph convolution. We first used the GCN model to conduct experiments on the network constructed based on sequence information, which is represented as GCN in the table. Then, we added mutual information on top of the sequence information to construct the network and conducted experiments using the GCN model, represented as MI-GCN. The results show that the inclusion of mutual information improved the prediction performance of the model.

Table 4. The effect of mutual information and RGC

| MI-RGC | AUPR | AUC | ACC |
|---|---|---|---|
| GCN | 0.3334 | 0.7934 | 0.8466 |
| MI-GCN | 0.4351 | 0.8546 | 0.9037 |
| RGC | 0.6495 | 0.9577 | 0.9275 |
| MI-RGC | 0.7072 | 0.9650 | 0.9300 |

To evaluate the performance of regional graph convolution, we used the regional graph convolutional model for feature learning on the network constructed from sequence information, represented as RGC. The results indicate that the prediction performance of the regional graph convolutional model is significantly higher than that of the GCN model. Finally, we considered both mutual information and regional graph convolution, represented as MI-RGC. The results demonstrate that the model obtained with this combination achieves the best prediction performance.

The experimental results are shown in Table 4. The experimental results show that adding the mutual information module to the graph convolution model significantly improved AUC and ACC scores. After replacing graph convolution with regional graph convolution, the model's AUC increased by 17%, and ACC increased by 9%. Although adding mutual information to the regional graph convolution also improved the experimental results, the improvement was limited. Our analysis suggests that, on one hand, the mutual information module contains a certain amount of redundant information; on the other hand, the scale of the mutual information data is also an important factor.

## Evaluation of region partitioning

This section primarily conducts experiments, comparisons, and analyses on different choices of region partitioning. Overall, the performance of the model after region partitioning has shown

Table 5. Results with a region length of 0.005

| Region | AUPR | AUC | ACC |
|---|---|---|---|
| $R = \{r_1\}$ | 0.4206 | 0.9463 | 0.9771 |
| $R = \{r_1, r_2\}$ | 0.4323 | 0.9518 | 0.9770 |
| $R = \{r_1, r_2, r_3\}$ | 0.4297 | 0.9523 | 0.9796 |
| $R = \{r_1, r_2, r_3, r_4\}$ | 0.4622 | 0.9634 | 0.9877 |
| $R = \{r_1, r_2, r_3, r_4, r_5\}$ | 0.4266 | 0.9504 | 0.9783 |
| $R = \{r_1, r_2, r_3, r_4, r_5\} \cup \{r_!\}$ | 0.3987 | 0.9057 | 0.9676 |

Table 6. Results with a region length of 0.01

| Region | AUPR | AUC | ACC |
|---|---|---|---|
| $R = \{r_1\}$ | 0.4266 | 0.9504 | 0.9783 |
| $R = \{r_1, r_2\}$ | 0.4472 | 0.9620 | 9735 |
| $R = \{r_1, r_2, r_3\}$ | 0.4359 | 0.9569 | 0.9728 |
| $R = \{r_1, r_2, r_3, r_4\}$ | 0.4317 | 0.9436 | 0.9681 |
| $R = \{r_1, r_2, r_3, r_4, r_5\}$ | 0.4176 | 0.9215 | 0.9522 |
| $R = \{r_1, r_2, r_3, r_4, r_5\} \cup \{r_!\}$ | 0.3967 | 0.8954 | 0.9253 |

significant improvement compared to the basic GCN model, as demonstrated in Tables 5 and 6.

For the region partitioning, we selected two interval lengths: 0.005 and 0.01 for the experiments. First, the neighbors are partitioned into regions based on the specified interval lengths: $R = \{r_1\}$, $R = \{r_1, r_2\}$,..., $R = \{r_1, r_2, r_3, ..., r_m\} \cup \{r_!\}$. Taking 0.01 as an example, $\{r_1\}$ represents the interval $[1, 0.99]$; $\{r_1, r_2\}$ represents the interval $[1, 0.99) \cup [0.99, 0.98]$, and so on. Notably, $\{r_!\}$ indicates that the remaining nodes not included in the first $m$ regions are uniformly assigned to this area; for example, $\{r_1, r_2\} \cup \{r_!\}$ represents $[1, 0.99) \cup [0.99, 0.98) \cup [0.98, 0]$. The experimental results indicate that whether we choose 0.005 or 0.01, the predictive performance of the model after partitioning is clearly superior to that of the GCN model before partitioning. Moreover, after adding the $\{r_!\}$ region, both partitioning scenarios showed a notable decline in performance, suggesting that information from low-similarity neighbors can negatively impact the contribution to the central node. Ultimately, we chose an interval length of 0.005 with $R = \{r_1, r_2, r_3, r_4\}$ as the final output result.

The choice of region length was based on the distribution of distances between each central node and its neighboring nodes, which is partially shown in Table 7. In Table 7, the first five phages infect the same host, while the last five infect another host. According to Table 7, using 0.005 as the region length, when the threshold is set at 0.96, the number of neighboring nodes for each node is already quite large. Additionally, in Tables 5 and 6, we also presented experimental results using different region lengths. The results indicate that using 0.005 as the region length is optimal. Moreover, as the threshold continues to decrease, the performance of the model starts to decline.

## Evaluation of parameter tuning

This section mainly discusses the settings of the basic parameters involved in the RGC model. This section mainly discusses the setting of basic parameters involved in the model. We adopted a grid tuning approach to adjust and analyze the six basic parameters of the model. This is because grid tuning is more intuitive compared to other tuning methods. As shown in Fig. 4. It should be noted that we first conducted a tuning process for dimensions and learning rate, so the maximum values shown in Fig. 4(a) are not the optimal results. The results obtained from the tuning process of layers and epochs are presented in Fig. 5.

# Case study

## Case study for mutual information

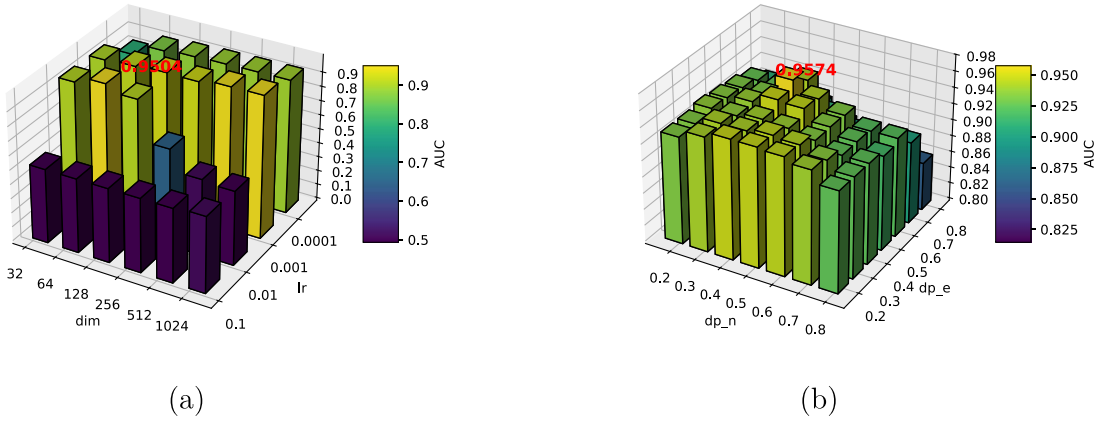
To demonstrate the reliability and effectiveness of mutual information more clearly, in addition to the relevant experiments we designed, we also conducted statistical analysis and case studies on the information content of mutual information. Through statistics, we obtained 16 052 pieces of mutual information, among which 12 894 pieces involve shared hosts. This means that in these 12 894 pieces of mutual information, the two phages involved in any mutual information infect the same bacteria in real data, which is precisely the information we need. Of course, the remaining mutual information that does not involve shared hosts also helps improve the performance of the experiments. Additionally, the number of phage species involved in these mutual information is 780. Taking the bacterium *Ruminococcus gnavus* as an example, as shown in Table 8, there are six phage species corresponding to this bacterium. Figure 6 illustrates the mutual information among these phages, indicating a strong association between them, which we attribute to their shared host.

## Case study for MI-RGC

To test the predictive performance of MI-RGC for new viruses and new host bacteria, we randomly selected two bacteria, *Klebsiellapneumoniae* and *Escherichiacoli*, as well as two phages,

Table 7. Partial presentation table of the distribution of neighboring nodes of 10 central nodes within the region

| Top | Phage | [1, 0.995) | [1, 0.99) | [1, 0.985) | [1, 0.98) | [1, 0.975) | [1, 0.97) | [1, 0.965) | [1, 0.96) |
|---|---|---|---|---|---|---|---|---|---|
| 1 | *Enterobacteria phage H19J* | 2 | 4 | 32 | 69 | 97 | 131 | 167 | 196 |
| 2 | *EnterobacteriaphageP-EibC* | 1 | 2 | 13 | 39 | 69 | 103 | 145 | 199 |
| 3 | *Enterobacteria phage P-EibD* | 1 | 2 | 12 | 15 | 27 | 60 | 82 | 116 |
| 4 | *EnterobacteriaphageP-EibE* | 2 | 3 | 18 | 53 | 83 | 112 | 141 | 178 |
| 5 | *Enterobacterial phage P-EibA* | 1 | 4 | 22 | 39 | 97 | 131 | 169 | 187 |
| 6 | *Lactococcus phage bIL311* | 31 | 61 | 106 | 167 | 255 | 328 | 378 | 428 |
| 7 | *Lactococcus phage bIL285* | 26 | 64 | 133 | 174 | 242 | 337 | 391 | 436 |
| 8 | *Lactococcus phage bIL286* | 25 | 66 | 130 | 170 | 278 | 349 | 400 | 446 |
| 9 | *Lactococcus phage bIL310* | 30 | 100 | 144 | 184 | 216 | 339 | 449 | 525 |
| 10 | *Lactococcus phage Tuc2009* | 38 | 93 | 135 | 208 | 297 | 340 | 449 | 561 |




(a) (b)

Figure 4. Experimental results based on grid tuning.

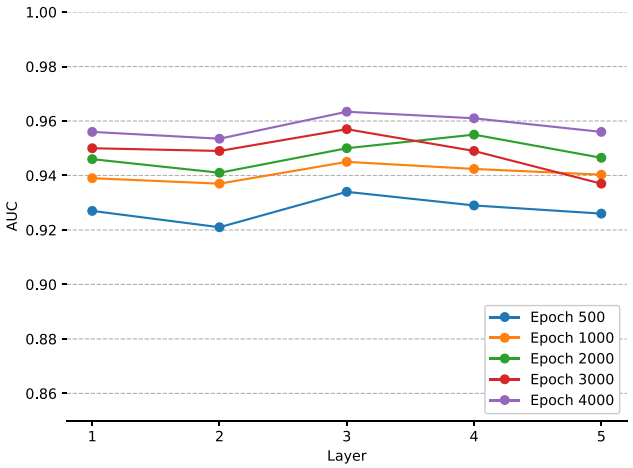



Figure 5. Visualization of the performance by layer and epoch.

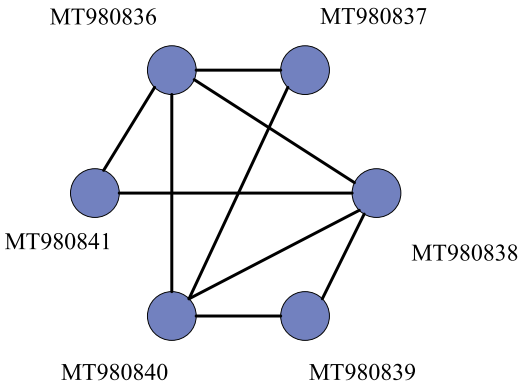



Figure 6. Visualization of the mutual information among the phages infecting *Ruminococcusgnavus*.

Table 8. The names and corresponding ACCESSION of the phages infecting *Ruminococcusgnavus*

| True-relevance | ACCESSION |
|---|---|
| *RuminococcusphagePHIsRg507T2-2* | MT980836 |
| *Ruminococcus phage PHIsRg507T2-3* | MT980837 |
| *Ruminococcus phage PHIsRM10* | MT980841 |
| *Ruminococcus phage PHIsRg519T2* | MT980838 |
| *Ruminococcus phage PHIsRgPS-6* | MT980839 |
| *Ruminococcus phage PHIsRgIBDN1* | MT980840 |

*phage* R18C and *Escherichia* 1720*a*-02. We removed all the association information of these four entities. We conducted experiments separately, and the results are shown in Tables 9, 10, and 11. The experimental results indicate that MI-RGC performed well in predicting new viruses and their corresponding new host bacteria.

## Discussion

In this study, we propose a novel prediction model (MI-RGC) for host prediction of novel viruses. Considering the characteristic of phages depending on host environments for survival, we leverage the concept of mutual information from information theory to extract effective information between phages from metagenomic data. Based on the extracted mutual information and existing sequence information, we construct heterogeneous networks: one based on mutual information and another based on sequence information. During the network embedding learning process, to better quantify the contributions of different neighboring nodes to the central node and avoid overfitting, we partition the neighboring nodes into different regions and use a region-level attention mechanism to learn the contributions of these regions. Finally, we use graph convolutional algorithms for information aggregation. When aggregating neighboring nodes within the same region, the nodes in that region share weights and parameters, which are not shared among different regions. The model then uses an attention mechanism to aggregate the embeddings of the two heterogeneous graphs, obtaining the final representation and using a predictor for prediction.

Experimental results show that the MI-RGC model outperforms state-of-the-art models in prediction performance. Additionally, we analyze the quality of the generated mutual information, demonstrating that the mutual information extracted from the sample environment is effective for phage host prediction tasks. Regarding the Region Graph Convolution (RGC) model, we also conducted experiments on different datasets, and the results indicate that RGC outperforms the latest model strategies in this task.

Table 9. The top20 results of predictions for *Klebsiellapneumoniae*

| Top | Phage | PMID | Top | Phage | PMID |
|---|---|---|---|---|---|
| 1 | *Klebsiella phage RAD2* | 34431721 | 11 | *Drulisvirus minorna* | N/A |
| 2 | *Klebsiella phage vB_KpnS-ZX4* | 26008965 | 12 | *Taipeivirus menlow* | 31023815 |
| 3 | *Klebsiella phage vB_1086* | 35985450 | 13 | *Delmidovirus copri* | N/A |
| 4 | *Klebsiella phage ST11-VIM1PHIs8.2* | 31752386 | 14 | *Webervirus mezzogao* | 29122857 |
| 5 | *Klebsiella phage vB_KpnM_17-11* | 35865823 | 15 | *Webervirus sweeny* | 31558643 |
| 6 | *Klebsiella phage vB_KleM KB2* | 37330608 | 16 | *Diorhovirus copri* | N/A |
| 7 | *Klebsiella phage vB_KpnP-Bp5* | 33631221 | 17 | *Webervirus sin4* | 31558644 |
| 8 | *Lastavirus sopranoga* | 29122857 | 18 | *Efquatrovirus SHEF4* | N/A |
| 9 | *Pylasvirus pylas* | 31727721 | 19 | *Drulisvirus altogao* | 29122857 |
| 10 | *Taipeivirus may* | 31072899 | 20 | *Yonseivirus seifer* | 31727722 |

Table 10. The top20 results of predictions for *Escherichiacoli*

| Top | Phage | PMID | Top | Phage | PMID |
|---|---|---|---|---|---|
| 1 | *Escherichia phage vB_EcoM_C2 — 3* | 34762992 | 11 | *Enterobacteria phage-P4* | 7483254 |
| 2 | *EscherichiaphagevB_EcoM-P10* | 34966369 | 12 | *Tequatrovirus-T4* | 26081634 |
| 3 | *Escherichia phage vB_EcoP-ZX5* | 36558779 | 13 | *Traversvirus II* | 12813092 |
| 4 | *EscherichiaphagevB_EcoM-Alf5* | 28522702 | 14 | *Goslarvirusgoslar* | 31109012 |
| 5 | *Escherichia phage vB_EcoS-PJ16* | 37632591 | 15 | *Inovirus M13* | 5257006 |
| 6 | *Escherichia coli phage PHIsStx2k* | 38078984 | 16 | *Kuttervirus SenALZ1* | N/A |
| 7 | *Escherichia phage vB_EcoM-RPN242* | 35598209 | 17 | *Escherichia phage OSYSP* | 38182094 |
| 8 | *Escherichia phage TL-2011b* | 22403614 | 18 | *Kayfunavirus SH4* | N/A |
| 9 | *Escherichia phage Schickermooser* | 31109012 | 19 | *Jahgtovirus intestinalis* | N/A |
| 10 | *Escherichia phage SRT7* | 30762120 | 20 | *Warwickvirus tunus* | 32899836 |

Table 11. The top5 results of predictions for *phageR18C* and *Escherichia1720a-02*

| Phage | Host-predicting | Evidence(PMID) |
|---|---|---|
| *phage R18C* | *Escherichia coli* | 31641840 |
| | *Citrobacterrodentium* | 31641840 |
| | *Erwinia amylovora* | N/A |
| | *Shigella sonnei* | 31641840 |
| | *Citrobacter koser* | N/A |
| *Escherichia 1720a-02* | *Citrobacter rodentium* | 34929548 |
| | *Escherichia sp.* | N/A |
| | *Citrobacter freundii* | N/A |
| | *Escherichia coli* | 32761142 |
| | *Citrobacter koseri* | N/A |

**Key Points**

- We annotated the state information of phages in each sample from metagenomic data, treated the state information as random variables for mutual information calculation, and ultimately obtained the mutual information of phages.
- We constructed a heterogeneous network using the mutual information of phages and known interactions, which was used as a feature augmentation module.
- We developed a Region Graph Convolution (RGC) model, which divides the network constructed based on distance metrics into different regions and learns the contributions of neighbors in different regions to the central node through a region-level attention mechanism.
- The design and prediction strategy of the MI-RGC model is significantly effective in PHIs prediction, and the end-to-end model is more conducive to applying this model in other fields.

## Acknowledgments

We express our deepest gratitude to the authors of all the code and data cited in this paper, especially Feiyue Sun and Jianqiang Sun for their GCNAT method, which greatly inspired our work. We also thank the reviewers for their helpful and constructive comments.

Conflict of Interest: None declared.

## Funding

This work was partially supported by the National Natural Science Foundation of China (62372205, 62472192, and 61932008), the National Language Commission Key Research Project (ZDI145-56), the Fundamental Research Funds for Central Universities (KJ02502022-0450), the Natural Science Foundation of Hubei Province of China (2022CFB289), and the Self-determined Research Funds of CCNU from the Colleges' Basic Research and Operation of MOE (CCNU24JC032).

## Data availability

The metagenomic datasets are labeled as PRJNA422434 and PRJNA834801, comprising 370 and 725 human gut samples, respectively. Related code, dataset splits, and trained models can be found at: https://github.com/Ankang-Wei/MI-RGC.

## Author contributions

**Ankang Wei:** contributed to the methodological ideas of MI-RGC, conducted experimental data analysis, visualized the results, and drafted the manuscript. **Zhen Xiao:** contributed to data organization and analysis, and drafted the manuscript. **Lingling Fu:** contributed to the organization and analysis of real data, and drafted the manuscript. **Weizhong Zhao:** contributed to the methodological ideas of MI-RGC and drafted the manuscript. **Xingpeng Jiang:** contributed to the methodological ideas, biological insights, and interpretations of MI-RGC, and drafted the manuscript. All authors have read and approved the final manuscript.

## References


1. Malik DJ, Sokolov IJ, Vinner GK. *et al.* Formulation, stabilisation and encapsulation of bacteriophage for phage therapy. *Adv Colloid Interface Sci* 2017;**249**:100–33. https://doi.org/10.1016/j.cis.2017.05.014.
2. Saha M, Sarkar A. Review on multiple facets of drug resistance: a rising challenge in the 21st century. *J Xenobiot* 2021;**11**:197–214. https://doi.org/10.3390/jox11040013.
3. Johansson MHK, Bortolaia V, Tansirichaiya S. *et al.* Detection of mobile genetic elements associated with antibiotic resistance in Salmonella enterica using a newly developed web tool: MobileElementFinder. *J Antimicrob Chemother* 2021;**76**:101–9. https://doi.org/10.1093/jac/dkaa390.
4. Rhoads DD, Wolcott RD, Kuskowski MA. *et al.* Bacteriophage therapy of venous leg ulcers in humans: Results of a phase i safety trial. *J Wound Care* 2009;**18**:237–43. https://doi.org/10.12968/jowc.2009.18.6.42801.
5. Federici S, Kredo-Russo S, Valdés-Mas R. *et al.* Targeted suppression of human IBD-associated gut microbiota commensals by phage consortia for treatment of intestinal inflammation. *Cell* 2022;**185**:2879–2898.e24. https://doi.org/10.1016/j.cell.2022.07.003.
6. Tetz G, Brown SM, Hao Y. *et al.* Type 1 diabetes: an association between autoimmunity, the dynamics of gut amyloid-producing E. coli and their phages. *Sci Rep* 2019;**9**:9685. https://doi.org/10.1038/s41598-019-46087-x.
7. Nie W, Qiu T, Wei Y. *et al.* Advances in phage–host interaction prediction: in silico method enhances the development of phage therapies. *Brief Bioinform* 2024;**25**:bbae117. https://doi.org/10.1093/bib/bbae117.
8. Bajiya N, Dhall A, Aggarwal S. *et al.* Advances in the field of phage-based therapy with special emphasis on computational resources. *Brief Bioinform* 2023;**24**:bbac574. https://doi.org/10.1093/bib/bbac574.
9. Edwards RA, McNair K, Faust K. *et al.* Computational approaches to predict bacteriophage–host relationships. *FEMS Microbiol Rev* 2016;**40**:258–72. https://doi.org/10.1093/femsre/fuv048.
10. Coclet C, Roux S. Global overview and major challenges of host prediction methods for uncultivated phages. *Curr Opin Virol* 2021;**49**:117–26. https://doi.org/10.1016/j.coviro.2021.05.003.
11. Krysiak-Baltyn K, Martin GJO, Stickland AD. *et al.* Computational models of populations of bacteria and lytic phage. *Crit Rev Microbiol* 2016;**42**:942–68. https://doi.org/10.3109/1040841X.2015.1114466.
12. Versoza CJ, Pfeifer SP. Computational prediction of bacteriophage host ranges. *Microorganisms* 2022;**10**:149. https://doi.org/10.3390/microorganisms10010149.
13. Ahlgren NA, Ren J, Yang Young L. *et al.* Alignment-free oligonucleotide frequency dissimilarity measure improves prediction of hosts from metagenomically-derived viral sequences. *Nucleic Acids Res* 2017;**45**:39–53. https://doi.org/10.1093/nar/gkw1002.
14. Congyu L, Zhang Z, Cai Z. *et al.* Prokaryotic virus host predictor: a Gaussian model for host prediction of prokaryotic viruses in metagenomics. *BMC Biol* 2021;**19**:1–11.
15. Ruohan W, Xianglilan Z, Jianping W. *et al.* DeepHost: phage host prediction with convolutional neural network. *Brief Bioinform* 2022;**23**:bbab385. https://doi.org/10.1093/bib/bbab385.
16. Shang J, Sun Y. Predicting the hosts of prokaryotic viruses using GCN-based semi-supervised learning. *BMC Biol* 2021;**19**:1–15. https://doi.org/10.1186/s12915-021-01180-4.
17. Shang J, Sun Y. CHERRY: a computational method for accurate prediction of virus–prokaryotic interactions using a graph encoder–decoder model. *Brief Bioinform* 2022;**23**:bbac182. https://doi.org/10.1093/bib/bbac182.
18. Li M, Zhang W. PHIAF: prediction of phage-host interactions with GAN-based data augmentation and sequence-based feature fusion. *Brief Bioinform* 2022;**23**:bbab348. https://doi.org/10.1093/bib/bbab348.
19. Boeckaerts D, Stock M, Criel B. *et al.* Predicting bacteriophage hosts based on sequences of annotated receptor-binding proteins. *Sci Rep* 2021;**11**:1467. https://doi.org/10.1038/s41598-021-81063-4.
20. Coutinho FH, Zaragoza-Solas A, López-Pérez M. *et al.* RaFAH: host prediction for viruses of bacteria and archaea based on protein content. *Patterns* 2021;**2**:100274. https://doi.org/10.1016/j.patter.2021.100274.
21. Amgarten D, Iha BKV, Piroupo CM. *et al.* vHulk, a new tool for bacteriophage host prediction based on annotated genomic features and neural networks. *PHAGE* 2022;**3**:204–12. https://doi.org/10.1089/phage.2021.0016.
22. Galiez C, Siebert M, Enault F. *et al.* WISH: who is the host? Predicting prokaryotic hosts from metagenomic phage contigs. *Bioinformatics* 2017;**33**:3113–4. https://doi.org/10.1093/bioinformatics/btx383.
23. Zhou F, Gan R, Zhang F. *et al.* PHISDetector: a tool to detect diverse in silico phage–host interaction signals for virome studies. *Genom Proteom Bioinform* 2022;**20**:508–23. https://doi.org/10.1016/j.gpb.2022.02.003.
24. Tang T, Hou S, Fuhrman JA. *et al.* Phage–bacterial contig association prediction with a convolutional neural network. *Bioinformatics* 2022;**38**:i45–52. https://doi.org/10.1093/bioinformatics/btac239.
25. Tan J, Fang Z, Shufang W. *et al.* HoPhage: an ab initio tool for identifying hosts of phage fragments from metaviromes. *Bioinformatics* 2022;**38**:543–5. https://doi.org/10.1093/bioinformatics/btab585.
26. Yeh T-K, Jean S-S, Lee Y-L. *et al.* Bacteriophages and phage-delivered CRISPR-CAS system as antibacterial therapy. *Int J Antimicrob Agents* 2022;**59**:106475. https://doi.org/10.1016/j.ijantimicag.2021.106475.
27. Takeuchi I, Osada K, Azam AH. *et al.* The presence of two receptor-binding proteins contributes to the wide host range of

staphylococcal twort-like phages. *Appl Environ Microbiol* 2016;**82**:5763–74. https://doi.org/10.1128/AEM.01385-16.
28. Silva JB, Storms Z, Sauvageau D. Host receptors for bacteriophage adsorption. *FEMS Microbiol Lett* 2016;**363**:01. https://doi.org/10.1093/femsle/fnw002.
29. León-Félix J, Villicaña C. The impact of quorum sensing on the modulation of phage-host interactions. *J Bacteriol* 2021;**203**:10–1128. https://doi.org/10.1128/JB.00687-20.
30. Qin J, Li Y, Cai Z. *et al.* A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature* 2012;**490**:55–60. https://doi.org/10.1038/nature11450.
31. Chibani-Chennoufi S, Bruttin A, Dillmann M-L. *et al.* Phage-host interaction: an ecological perspective. *J Bacteriol* 2004;**186**:3677–86. https://doi.org/10.1128/JB.186.12.3677-3686.2004.
32. Neelakanta P, Chatterjee S, Pappusetty D. *et al.* Information-theoretic algorithms in bioinformatics and bio-/medical-imaging: a review. In: *2011 International conference on recent trends in information technology (ICRTIT)*, IEEE. 2011, pp. 183–8.
33. McIver LJ, Abu-Ali G, Franzosa EA. *et al.* bioBakery: a meta'omic analysis environment. *Bioinformatics* 2018;**34**:1235–7. https://doi.org/10.1093/bioinformatics/btx754.
34. Andrews S. FastQC: a quality control tool for high throughput sequence data. *Babraham Bioinform;* 2010. Available at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc.
35. Ho SFS, Wheeler NE, Millard AD. *et al.* Gauge your phage: benchmarking of bacteriophage identification tools in metagenomic sequencing data. *Microbiome* 2023;**11**:84. https://doi.org/10.1186/s40168-023-01533-x.
36. Jennifer L, Breitwieser FP, Thielen P. *et al.* Bracken: Estimating species abundance in metagenomics data. *PeerJ Comput Sci* 2017;**3**:e104.
37. Gomes M, Hamer R, Reinert G. *et al.* Mutual information and variants for protein domain-domain contact prediction. *BMC Res Notes* 2012;**5**:1–14. https://doi.org/10.1186/1756-0500-5-472.
38. Li M, Wang Y, Li F. *et al.* A deep learning-based method for identification of bacteriophage-host interaction. *IEEE/ACM Trans Comput Biol Bioinform* 2020;**18**:1801–10. https://doi.org/10.1109/TCBB.2020.3017386.
39. Sun F, Sun J, Zhao Q. A deep learning method for predicting metabolite–disease associations via graph neural network. *Brief Bioinform* 2022;**23:bbac266**. https://doi.org/10.1093/bib/bbac266.
40. Zhang Y-z, Liu Y, Bai Z. *et al.* Zero-shot-capable identification of phage–host relationships with whole-genome sequence representation by contrastive learning. *Brief Bioinform* 2023;**24**:bbad239. https://doi.org/10.1093/bib/bbad239.

© The Author(s) 2024. Published by Oxford University Press. This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (https://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com
*Briefings in Bioinformatics*, 2025, **26(1)**, bbae672
https://doi.org/10.1093/bib/bbae672
Problem Solving Protocol