Briefings in Bioinformatics, 2024, **25(6)**, bbae484
https://doi.org/10.1093/bib/bbae484
Problem Solving Protocol

DeepPBI-KG: a deep learning method for the prediction of phage-bacteria interactions based on key genes

Tongqing Wei[1,‡], Chenqi Lu[1,‡], Hanxiao Du[1], Qianru Yang[1], Xin Qi[2], Yankun Liu[2], Yi Zhang[3], Chen Chen[3], Yutong Li[1], Yuanhao Tang[1], Wen-Hong Zhang[1,2,3,*], Xu Tao[1,2,3,*], Ning Jiang[1,2,3,*]

[1]State Key Laboratory of Genetic Engineering, School of Life Sciences, Fudan University, No. 2005 Songhu Road, Shanghai, 200433, China
[2]Shanghai Sci-Tech Inno Center for Infection & Immunity, No. 1688 Guoquan Bei Road, Shanghai, China
[3]Department of Infectious Diseases, Huashan Hospital, Shanghai Medical College, Fudan Univerisy, No. 12 Wulumuqi Zhong Road, Shanghai, China

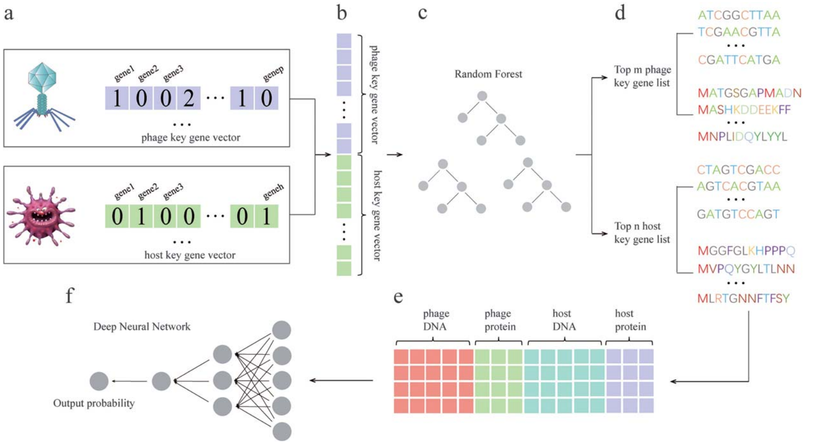
[*]Corresponding authors. Ning Jiang, State Key Laboratory of Genetic Engineering, School of Life Sciences, Fudan University, Shanghai, China. E-mail: ningjiang@fudan.edu.cn; Xu Tao, Department of Infectious Diseases, Huashan Hospital, Shanghai Medical College, Fudan University, Shanghai, China. E-mail: xutao@fudan.edu.cn; Wen-Hong Zhang, Department of Infectious Diseases, Huashan Hospital, Shanghai Medical College, Fudan University, Shanghai, China. E-mail: zhangwenhong@fudan.edu.cn

[‡]Tongqing Wei and Chenqi Lu contributed equally to this work.

Abstract

Phages, the natural predators of bacteria, were discovered more than 100 years ago. However, increasing antimicrobial resistance rates have revitalized phage research. Methods that are more time-consuming and efficient than wet-laboratory experiments are needed to help screen phages quickly for therapeutic use. Traditional computational methods usually ignore the fact that phage-bacteria interactions are achieved by key genes and proteins. Methods for intraspecific prediction are rare since almost all existing methods consider only interactions at the species and genus levels. Moreover, most strains in existing databases contain only partial genome information because whole-genome information for species is difficult to obtain. Here, we propose a new approach for interaction prediction by constructing new features from key genes and proteins via the application of K-means sampling to select high-quality negative samples for prediction. Finally, we develop DeepPBI-KG, a corresponding prediction tool based on feature selection and a deep neural network. The results show that the average area under the curve for prediction reached 0.93 for each strain, and the overall AUC and area under the precision-recall curve reached 0.89 and 0.92, respectively, on the independent test set; these values are greater than those of other existing prediction tools. The forward and reverse validation results indicate that key genes and key proteins regulate and influence the interaction, which supports the reliability of the model. In addition, intraspecific prediction experiments based on *Klebsiella pneumoniae* data demonstrate the potential applicability of DeepPBI-KG for intraspecific prediction. In summary, the feature engineering and interaction prediction approaches proposed in this study can effectively improve the robustness and stability of interaction prediction, can achieve high generalizability, and may provide new directions and insights for rapid phage screening for therapy.

Graphical Abstract



Keywords: phage-bacteria interaction; machine learning; deep learning; negative sample selection; receptor binding protein

Received: May 30, 2024. **Revised:** August 18, 2024. **Accepted:** September 13, 2024

Introduction

Currently, the global burden of treating bacterial infections remains significant, severely impacting human health and quality of life. Numerous reports indicate that bacterial infections are associated with the spread and progression of a wide range of diseases [1], especially bacterial infections in the lungs and gut [2, 3]. In prior research, antibiotics and antimicrobial peptides have been employed for the treatment of bacterial infections. However, as bacteria evolve and increasingly develop resistance to these treatments, phage therapy has gradually emerged as an option for treating bacterial infections, demonstrating strong vitality and therapeutic efficacy [4]. Selecting appropriate phages capable of killing bacteria through experimental methods is tedious and labour-intensive. Saving time and effort to select the required phages swiftly represents a significant challenge in this field of research currently.

Rapid advancements in computer technology and computational biology have made it feasible to screen phages through computational methods. Consequently, computational tools for predicting phage-bacterial interactions on a large scale have emerged [5]. At present, methods for predicting phage-bacteria interactions can be broadly categorized into three types: first, methods based on similar phages and bacteria, such as k-mer similarity analysis, basic local alignment search tool (BLAST) alignment, and various other methods that use alignment software to construct feature vectors; second, methods integrating various sequence features of phages and bacteria to construct negative samples for calculation; and third, methods involving the extraction of receptor-binding phage proteins (RBPs), specifically tail proteins, which are directly related to interactions, and computing their features at the DNA and protein levels. Recently, tools developed on the basis of these methods have included predicting phage-host interactions (PredPHI) [6], DeepHost [7], who is the host (WIsH) [8], VirHostMatcher [9], prokaryotic virus host predictor (PHP) [10], viral host unveiling kit (vHULK) [11], a Python package for predicting phage-bacteria interactions through the local k-mer strategy (PB-LKS) [12], and ViWrap [13]. vHULK constructs feature matrices through alignment methods and employs a multilayer perceptron model for multiclass prediction. PHP relies on k-mer similarity, VirHostMatcher utilizes oligonucleotide frequency similarity, WIsH trains Markov models, and PredPHI constructs DNA-protein features and a convolutional neural network for binary classification prediction. PB-LKS predicted interactions via a local k-mer-based strategy. ViWrap is a comprehensive pipeline platform that combines virus identification, annotation, genome classification, and host prediction functions.

However, computational methods for phage-bacteria interaction prediction have not yet been perfected, and their predictive performance currently reaches that of the state-of-the-art (SOTA) methods within only individual strains, lacking broad applicability and generalizability. The presence of substantial redundant information in whole-genome sequence (WGS) predictions indicates that there remains room for improvement in predictive performance. Phage-bacteria interactions are primarily mediated by interactions among key proteins and are regulated by certain critical genes [14, 15]. Therefore, whole-genome predictions are affected by redundant information. In this study, DeepPBI-KG, a novel feature construction method and predictive framework that enhances the interpretability of model predictability from the perspective of critical proteins and key genes, is introduced. First, a validation module based on key genes is designed to enhance the interpretability of the model, revealing its biological significance. Second, diverging from previously described approaches that locate the WGSs of phages and bacteria, in this study, the interactions are narrowed to selectively screen key genes directly related to the interaction [16]. We constructed our model framework on this basis. Third, we employed K-means to construct high-quality negative samples, creating a balanced dataset and enhancing the predictive performance. Benchmark tests indicate that DeepPBI-KG achieves robust predictive performance.

Materials and methods

Dataset

We extracted information from five public datasets to compile the data necessary for constructing the model, including data on phages, bacteria, and their interactions, which can be accessed from published datasets or widely used public databases such as PredPHI [6], PhageHosts [5], National Center of Biotechnology Information (NCBI) [17], Microbe Versus Phage (MVP) [18], and PHISDetector [19]. These datasets were combined to construct the dataset for this study. The data from the five sources were filtered, and duplicates were removed. Initially, phages and bacteria that were not reported in relevant literature or not recorded in the NCBI database were eliminated to ensure data reliability. Subsequently, incorrectly labelled phages and bacteria were filtered out, and the corresponding WGSs were extracted from the NCBI database. Third, phages and bacteria with overly short, incomplete, or inaccurately sequenced genomes were removed. Following these procedures, the remaining phage-bacteria interactions from the five published papers and public databases were merged, eliminating duplicates to yield 3647 interactions between 3513 phages and 395 bacteria as a training and test set, and 1230 interactions between 1182 phages and 166 bacteria as an independent and external test set. Additionally, the WGSs obtained from the NCBI database were annotated with Prokka [20], generating annotated GBK files that contained all coding sequence information for phage and bacteria. We extracted the protein sequences of the coding sequences (cdss) from the annotated GBK file and aligned them with the Conserved Domain Database via BLASTP to identify the primary function of these cdss, which included many hypothetical proteins [21], to extract the phage tail proteins and bacteria receptor protein to construct the RBP dataset. The organized RBP dataset was compared with the key genes selected by the model in subsequent studies.

Sampling methods

Interaction prediction is framed as a binary classification task to predict whether a phage interacted with a bacterium, where an interaction is denoted by 1 and no interaction is denoted by 0. Hence, positive samples are defined as phage-bacteria pairs that interact with each other. Given the scarcity of naturally interacting phage-bacteria pairs in the environment, researchers often focus on these interacting pairs. Consequently, the samples obtained are positive samples that exist in nature and have been experimentally verified. However, negative samples require artificial construction. In this study, samples (interaction or non-interaction phage-bacteria pairs) that did not appear among the positive samples after a certain random matching procedure were considered negative samples.

However, different sampling methods can impact the learning effectiveness of the model differently. Moreover, to achieve good model performance, the prediction trends of positive and negative samples should be consistent, and their quantities should be approximately equivalent to maintain sample balance. The number of negative samples generated by random matching significantly exceeds that of positive samples. Therefore, selecting the most representative batch of samples from a large pool of negative samples is a critical consideration. The K-means method was employed to choose high-quality negative samples [22] to enable the model to learn the distinctions between positive and negative samples better. All negative samples were subjected to K-means clustering, with the k value set to the number of positive samples, ensuring that the resulting negative samples were more diverse and representative. When selecting samples via K-means, we used Python's scikit-learn library and NumPy library for data preprocessing, including standardization, scale, and processing of missing data. We also constructed negative samples through random sampling and compared the training effects of the two sampling methods to highlight the advantages of K-means sampling.

Enrichment analysis based on key genes and selection of gene scoring threshold under RF

A key_gene feature vector was constructed, and the random forest (RF) model was selected for feature screening to investigate the key genes affecting phage-bacteria interactions. We designed an enrichment analysis verification method based on key_gene to increase the interpretability of the model and verify that phage-bacteria interactions are essentially interactions of key proteins in the genomic sequence. Specifically, each gene of the phage (or bacterium) was sorted in ascending order of feature importance predicted by the RF model, and a threshold was set to select the most important genes. The selection method for this threshold is similar to the principle of principal component analysis, ensuring that the sum of the feature importance of genes ranked above the threshold accounts for >90% of the total feature importance sum, meaning that the contribution of the top genes reaches 90%. Since typically only a few key features truly influence interactions, the computationally features with high weights facilitate biological interpretation of the model. For forward validation, Kyoto Encyclopedia of Genes and Genomes (KEGG) enrichment analysis and Gene Ontology (GO) enrichment analysis were conducted on the top genes filtered by the RF model to validate their relationship with phage-bacteria interactions. If relevant, for external data predictions, only the key gene sequences of the corresponding species needed to be extracted and features needed to be calculated for accurate model predictions. Moreover, applying these features eliminates the impact of redundant information on the basis of whole-genome data. Reverse validation involved observing the ordering of genes included in RBP dataset within the gene set sorted on the basis of their importance predicted by the RF model.

Model constructions

A binary predictive model for phage-bacteria interactions was established. Notably, the quantity of phage samples significantly exceeds that of bacteria, leading to scenarios where a single bacterium might interact with multiple phages. However, in 80% of cases, bacteria generally interact with only a small number of phages (no >10), meaning that only a few dozen bacteria interact with hundreds of phages. These interactions account for more than over half of the 3647 positive samples, resulting in considerable redundancy in the training set information. Therefore, the number of interacting phage samples per bacterium was capped at 10 by under sampling. This threshold was selected by plotting the distribution of bacterial sample numbers to find the inflexion point and based on the effectiveness of model training (Supplementary Fig. S1). For bacteria that interact with >10 phages, the following procedure was employed: K-means was used to select the 10 most representative interacting phages for inclusion in the training set, with the remaining samples allocated to the test set. After sample filtration, the training set consisted of 1211 samples for model training, whereas the test set comprised 2436 samples for model testing and evaluation. Training was conducted via a deep neural network (DNN) with five hidden layers and a binary cross-entropy loss function, with the best-performing model parameters selected through cross-validation. The model was evaluated in terms of precision, accuracy, F1 score, recall, area under the curve (AUC), and area under the precision-recall curve (AUPR).

Sequencing information and experimental validation of bacteriophage - *Klebsiella pneumoniae* interactions

Total DNA was extracted from the bacteriophage and *K. pneumoniae* (KP) isolates using the TIANAmp DNA Kit (TIANAmp, Tiangen Biotech, Tiangen, China) according to the manufacturer's recommendations. After second-strand DNA was synthesized, DNA libraries were constructed by DNA fragmentation, end repair, A-tail addition, adapter ligation, and PCR amplification. An Agilent 2100 Bioanalyzer (Agilent, Santa Clara, CA) was used for quality control of the DNA libraries. Whole-genome sequencing was performed on the NovaSeq platform (Illumina, San Diego, CA) by constructing a paired-end (PE) library with average insertion lengths of 400–500 base pairs (bp). The raw sequencing data (1 gb/sample) were preprocessed as follows: (1) adapter sequences were removed, (2) reads with >20 bp of low quality (Phred quality score <20) were removed, (3) reads with over 20 bp of ambiguous bases were removed, and (4) duplicated reads were removed. De novo assembly was performed using SPADES v3.11.1. The assembled genomic sequences were annotated via Prokka v1.14.6. The WGS data reported in this paper have been deposited in the Genome Warehouse at the National Genomics Data Center [23, 24], Beijing Institute of Genomics, Chinese Academy of Sciences/China National Center for Bioinformation, under an accession number that is publicly accessible at https://ngdc.cncb.ac.cn/gwh.

To determine the host range of each bacteriophage, the double agar overlay plaque assay was performed following a previously reported protocol [25]. Briefly, strains from the Kpn clinical strain library were grown to log phase and dispensed on LB agar. Each bacteriophage was prepared and quantified. The diluents were mixed with agar and dropped on the host bacterium, and allowed to solidify. Plaques were observed after 24 h of culture.

Results and discussion

Summary of data

To substantiate the diversity and broad-spectrum nature of the phages in the collected data, we adopted the approach described in a previous paper [26]. Figure 1A shows that phages with the ability to kill bacteria were abundant within phyla of Actinobacteria, Bacillota, and Pseudomonadota. Phages that infect Actinobacteria species, specifically species of the genus *Mycolicibacterium*, constituted 36% of the training and test sets (for detailed information on the training and test sets, refer to the **Dataset** section under **MATERIALS AND METHODS**). Owing to their large numbers, these phages are not displayed in Fig. 1A. However, we constructed an evolutionary tree comprising

1337 phages with the ability to kill *Mycolicibacterium* species, which further demonstrated the diverse and broad-spectrum nature of the phage data (Fig. 1B). Additionally, the number of phages significantly exceeds that of hosts, with a single host capable of interacting with multiple phages. A considerable distribution difference between phages and hosts (Fig. 1C and D) was observed among the 3647 interactions in the training and test sets.

Feature extraction

Numerous studies have indicated that DNA and protein sequences are crucial in the biological evolution of phages and their hosts [27]. We extracted features from the DNA and protein sequences of both phages and bacteria to acquire more valuable key insights, as outlined in Supplementary Table S1. The WGS lengths of phages and bacteria differ significantly, with phages typically spanning a few thousand bp and bacteria spanning hundreds of thousands of bp. Consequently, several sequence features unrelated to length were considered. These specific features were inspired primarily by characteristics involved in previous interaction studies, including DNA features, protein features, and some peptide features [28]. We calculated 90 features for the translated protein sequences and 133 features for the corresponding DNA sequences. The features were subsequently integrated as depicted in Fig. 2A: (a) Initially, each coding sequence of a phage (or bacterium) was designated cdss 1, cdss 2, ..., cdss m, where m signifies the number of coding sequences. The 133 DNA features and 90 protein features were marked as f 1, f 2, ..., f n, with n equating to 223. The next step involved calculating the maximum (max), minimum (min), average (mean), standard deviation (std), median, and variance (var) values for all cdss for the f 1 feature. (b) The max, min, mean, std, median, and var values for all cdss were calculated from f 2 to f n, resulting in the corresponding feature vectors. (c) The six vectors were concatenated to form a '6 × 223' dimensional vector, obtaining the DNA-protein feature carrier for a single phage (or bacterium).

Furthermore, a key gene auxiliary feature was constructed to aid in selecting critical genes associated with interactions. This feature was represented as a pseudo multihot vector, constructed as shown in Fig. 2B. First, a count was performed for the genes of all cdss of each phage (or bacterium). The gene and protein information for cdss was obtained from the GBK files. Specifically, a set of genes contained in all phages (bacteria) was denoted as $Gene_{item1}$, $Gene_{item2}$, ..., $Gene_{itemN}$, where item and N represented the phage (or bacterium) and the number of phages (or bacteria), respectively. A summary of the gene types across all phages (or bacteria) was subsequently performed to obtain the $Gene_{total}$ set. All the elements in the gene set were marked as gene1, gene2, ..., genek, where k indicates the number of genes. Next, the key gene feature vector was constructed, with k genes serving as the column dimensions of this vector. If a phage (or bacterial) cdss contains gene i, it was marked on the corresponding dimension for gene i in the key gene vector, with one occurrence marked as '1', two occurrences marked as '2', and so on. If a phage (or bacterium) lacked gene i, then the gene i dimension of the key_gene vector was marked as '0'.

Model framework

The specific model framework is shown in Fig. 2C: (a) constructed feature vectors of key_gene are required for both phages and bacteria; (b) the feature vectors of key_gene from phages and bacteria are concatenated to generate integrated features, where each sample represents a combined vector of a phage-bacteria pair; (c) positive and negative samples, integrated into combined features, are fed into an RF model for binary classification prediction, and the feature importance of each key genes is calculated; (d) genes with higher scores in phages and bacteria are selected on the basis of a predefined threshold, and the DNA and protein sequences corresponding to the top genes are extracted; (e) The DNA-protein feature corresponding to the top genes is computed and concatenated into a high-dimensional vector of dimensions '6 × 223 × 2', which serves as the final input feature for the model; and (f) a DNN model with five hidden layers is employed to predict the interaction probability.

Performance evaluation

Initially, training was conducted on a total of 3647 positive samples and an equal number of negative samples selected via K-means, using the RF model with key_gene features as input. On the basis of the feature importance predicted by RF, the top gene sets of phages and bacteria were selected. The cdss of the phages and bacteria were subsequently filtered according to the top gene set provided by the RF model, and the DNA-protein features of the phages and bacteria in the training set were subsequently calculated. Finally, the DNN model was employed for training, and its predictive performance was observed on the test set. The prediction of interactions on the basis of the DNA-protein features of key genes shows promising results. The essence of phage-bacteria interactions lies in protein–protein interactions, indicating that DNA-protein features, especially the DNA-protein features of key genes, can be used to effectively predict phage-bacteria interactions. The DeepPBI-KG model achieved a prediction accuracy of 0.87 and an AUC of 0.92 (Fig. 3A). It was also evident that the predictive performance of K-means sampling surpasses that of random sampling, which aligns with expectations.

Validation based on key genes

In the RF model, the feature importance of key_gene features indicates that among the multitude of genes possessed by phages and bacteria, the weight coefficients of most genes are close to zero, with only a minority of feature coefficients being relatively significant. These findings suggest that these few significant features truly influence the interaction between phages and bacteria and that such interactions can be explained by several notable genes and proteins (Fig. 3B). Feature importance thresholds were defined on the basis of the degree of contribution, setting the threshold for phages at 0.0001 and for hosts at 0.00004 (for detailed methodology on the selection of thresholds, refer to the Enrichment analysis based on key genes and selecting of gene scoring threshold under RF section under **MATERIALS AND METHODS**).

Therefore, genes with weight coefficients above the threshold were defined as key_gene features and were selected for forward validation. The phage genes were enriched primarily in DNA recombination, defence response to bacterium and DNA integration, with the main enriched pathways being DNA replication protein and DNA repair and recombination proteins (Fig. 3C, Supplementary Fig. S2A). The host genes were enriched predominantly in DNA integration, *N*-acetyltransferase activity, and pathways involved in nucleotide excision repair and the cell cycle—Caulobacter (Fig. 3D, Supplementary Fig. S2B). The enriched functions and pathways are related to phage-host infection, where phages invade the host cell and insert their genes

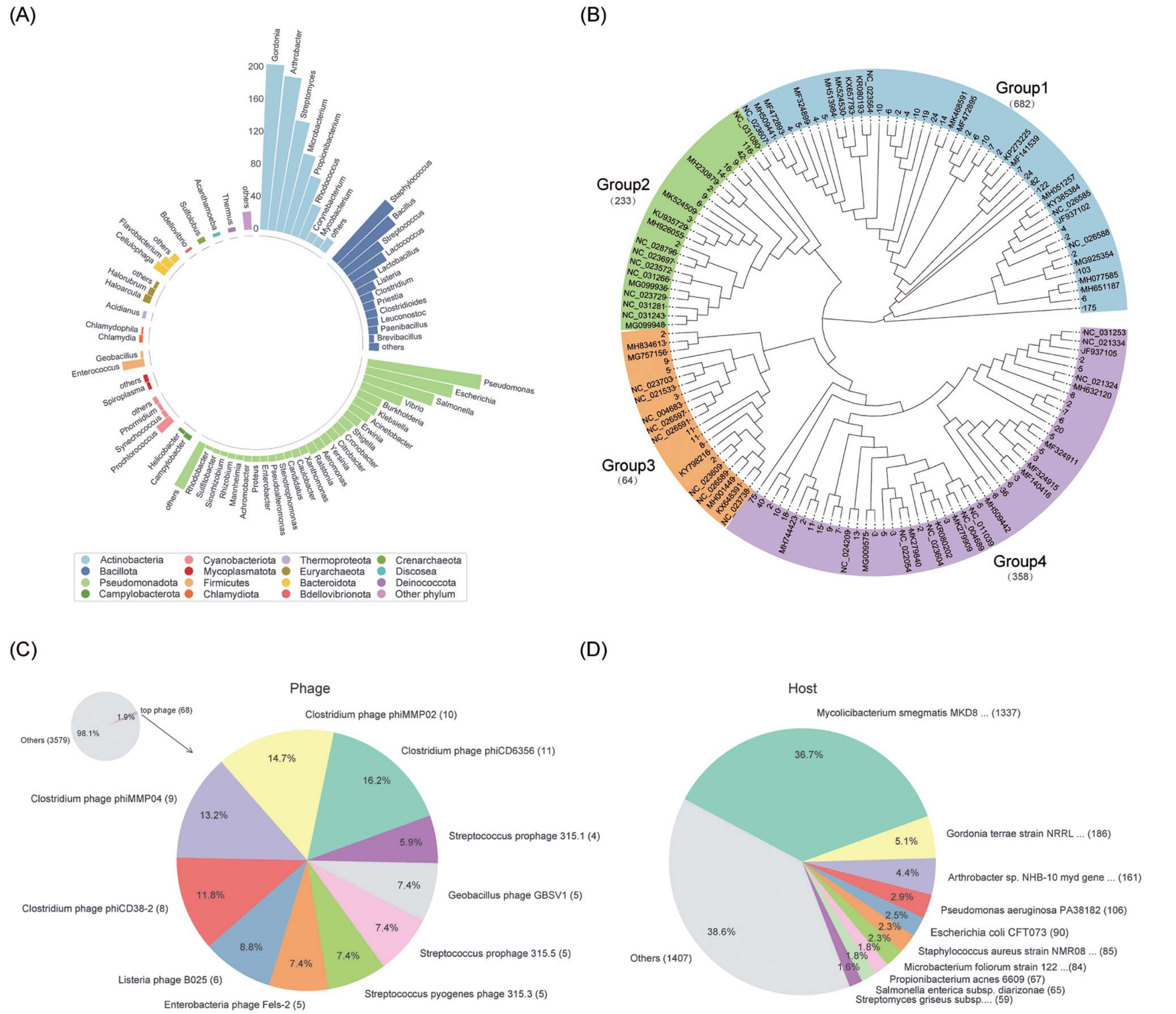


Figure 1. Summary statistics of the data. (A) Phage infection corresponding to the host of each phylum in the proportion toroidal histogram. The phage was statistically classified according to the name of the host species corresponding to the infection and displayed according to the information of the host phylum, wherein hosts with fewer than three members under each phylum were merged into others, and phyla containing too few host species were merged into other phyla. (B) Evolutionary tree of 1337 phage infected with *Mycolicibacterium*. Owing to the excessive number of phages involved and the complexity of the evolutionary tree, the evolutionary tree was pruned, and the subtree containing too many phages was merged. The number shown in the figure represents the number of phages contained in the subtree with the node as the root; the node showing the specific phage ID indicates that there is no subtree under the node. The evolutionary tree of *Mycolicibacterium*-infecting phages was divided into four clusters, and the number of phages contained in each cluster was labelled. (C) Distribution of phage in the interactions dataset. (D) Distribution of host in the interactions dataset.

into the bacterial DNA for replication, thereby killing the host (the related genes are gene = hin_1/gene = intA_5/gene = intS/gene = rdgC/gene = ligD/gene = recU/gene = rsxC_1 etc. The specific gene list can be found in the AVAILABILITY OF DATA AND MATERIAL Github project link).

Performance of the RBP and whole genome dataset

Phage-host interaction prediction is based on the results of plaque assays. Thus, it encompasses the recognition of the host by the phage (tail protein) and the infection of the host (DNA replication). We constructed both whole-genome and RBP datasets to illustrate this point. DeepPBI-KG was trained on key gene sequences selected by the RF model. In contrast, the whole-genome dataset involves training using the complete genome sequences of both phages and bacteria. The RBP dataset was compiled on the basis of previous research in which phage tail proteins and host interaction-related receptors were extracted [16, 29, 30] and subsequently used for training. Negative samples for all three datasets were collected via K-means, and DNA-protein features were computed, followed by training with the DNN model.

Compared with DeepPBI-KG, DeepPBI-WGS utilizes the full sequence information of both phages and bacteria, whereas the key_gene dataset contains only critical genes selected by the model, extracting key proteins directly related to interactions. Although the variety of proteins was reduced, the prediction accuracy and metrics (such as AUC) were slightly greater than those for the entire genome. This suggests that interactions were influenced only by key proteins and those interactions were interpreted and predicted using only key proteins (Table 1). Notably,

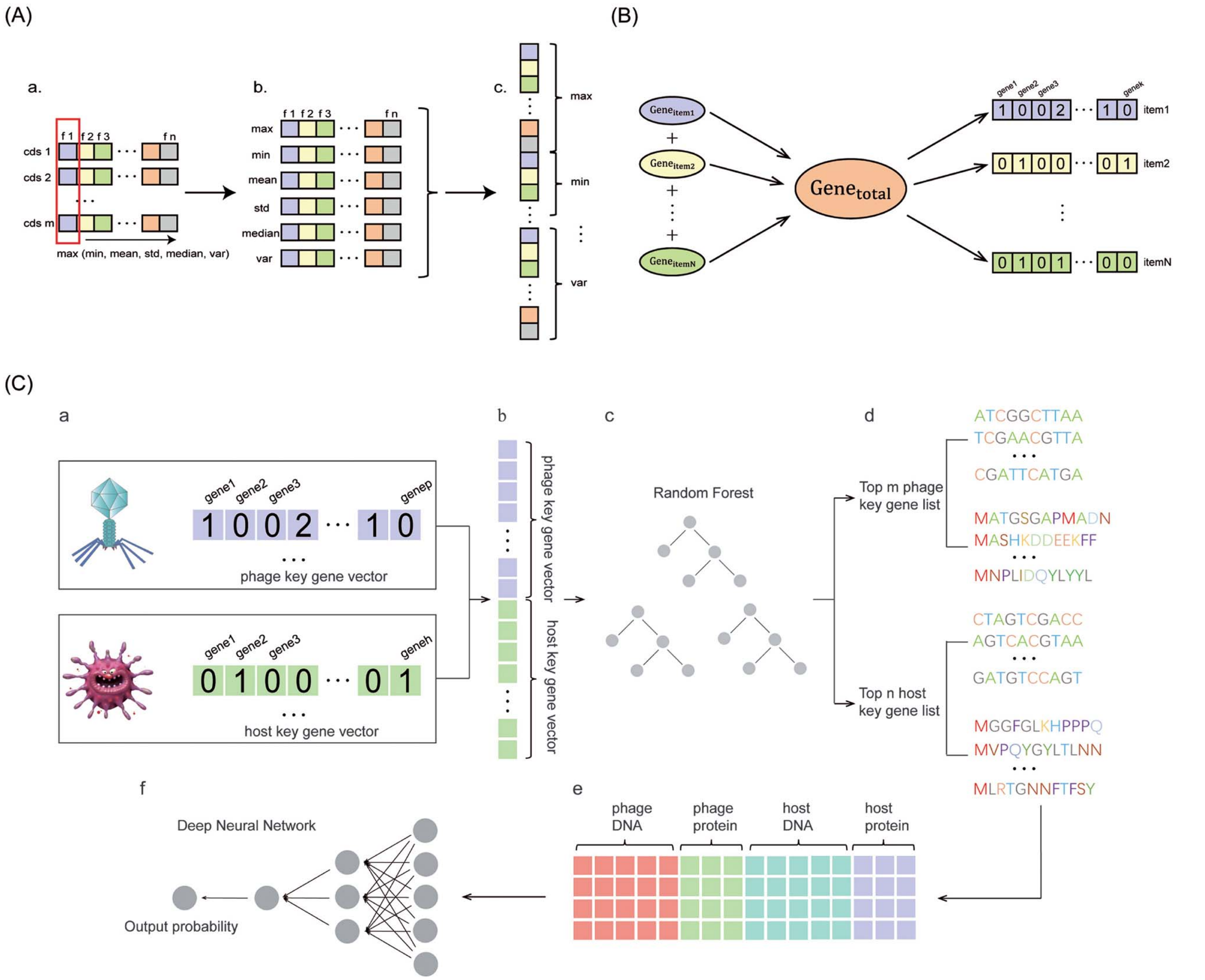


Figure 2. Feature construction and model framework. (A) DNA-protein feature construction process. (B) Key_gene feature construction process. (C) Model architecture flowchart.

K-means under DeepPBI-WGS was also better than random selection (Supplementary Table S2).

Similarly, compared with DeepPBI-KG, DeepPBI-RBP represents a dataset that encompasses manually curated key genes that influence interactions. Although the diversity of proteins was lower than that for the whole genome, the prediction accuracy and AUC of DeepPBI-RBP still fell short of those of DeepPBI-KG. This discrepancy suggests that some of the key genes identified through machine selection were more likely to have a genuine impact on phage-bacteria interactions than those selected manually (Table 1).

Comparison with state-of-the-art methods

We selected several representative predictive tools for comparison with our method, including PredPHI, DeepHost, WIsH, VirHostMatcher, PHP, vHULK, and PB-LKS. PB-LKS provides two prediction methods: RF and XGBoost. Figure 4A and B depicts the comparison of various methods in terms of the AUC and AUPR on the external test data, with the solid line representing the DeepPBI-KG method. Our method achieved a very high AUC value (AUC = 0.8917), surpassing all the other methods. Moreover, DeepPBI-KG ranked first among all methods in terms of AUPR (AUPR = 0.9218). Notably, DeepHost (AUC = 0.8253, AUPR = 0.8814) and PB-LKS (AUC = 0.8914, AUPR = 0.8999) exhibited greater predictive performance and composite rankings across multiple tools. The distinction between DeepHost and traditional predictive tools lies in its utilization of a novel DNA sequence encoding approach based on base matching and relative positioning without relying on the physical and chemical properties of the genome sequence during prediction. In contrast, PB-LKS is predicted via a local k-mer strategy. This process begins by fragmenting the whole genome of both the phage and the bacteria, followed by encoding these fragments. The most similar k-mer fragments between the phage and the bacteria are subsequently calculated. The difference in the features of the most similar fragment is then calculated as a measure of the divergence between them. Finally, a machine learning model is employed to learn the feature vector of the differences. Among them, VirHostMatcher, vHULK, DeepHost, and PB-LKS outperformed the other three tools in terms of AUC and AUPR. By comparing other evaluation metrics for DeepPBI-KG with those for VirHostMatcher, vHULK, DeepHost, and PB-LKS, the overall performance of the model was examined. The results indicate that DeepPBI-KG was the most stable across all other evaluation metrics, except for PB-LKS, which performs similarly to DeepPBI-KG. The other three tools exhibited a skewed distribution, tending towards predicting samples as negative (Fig. 4C). When

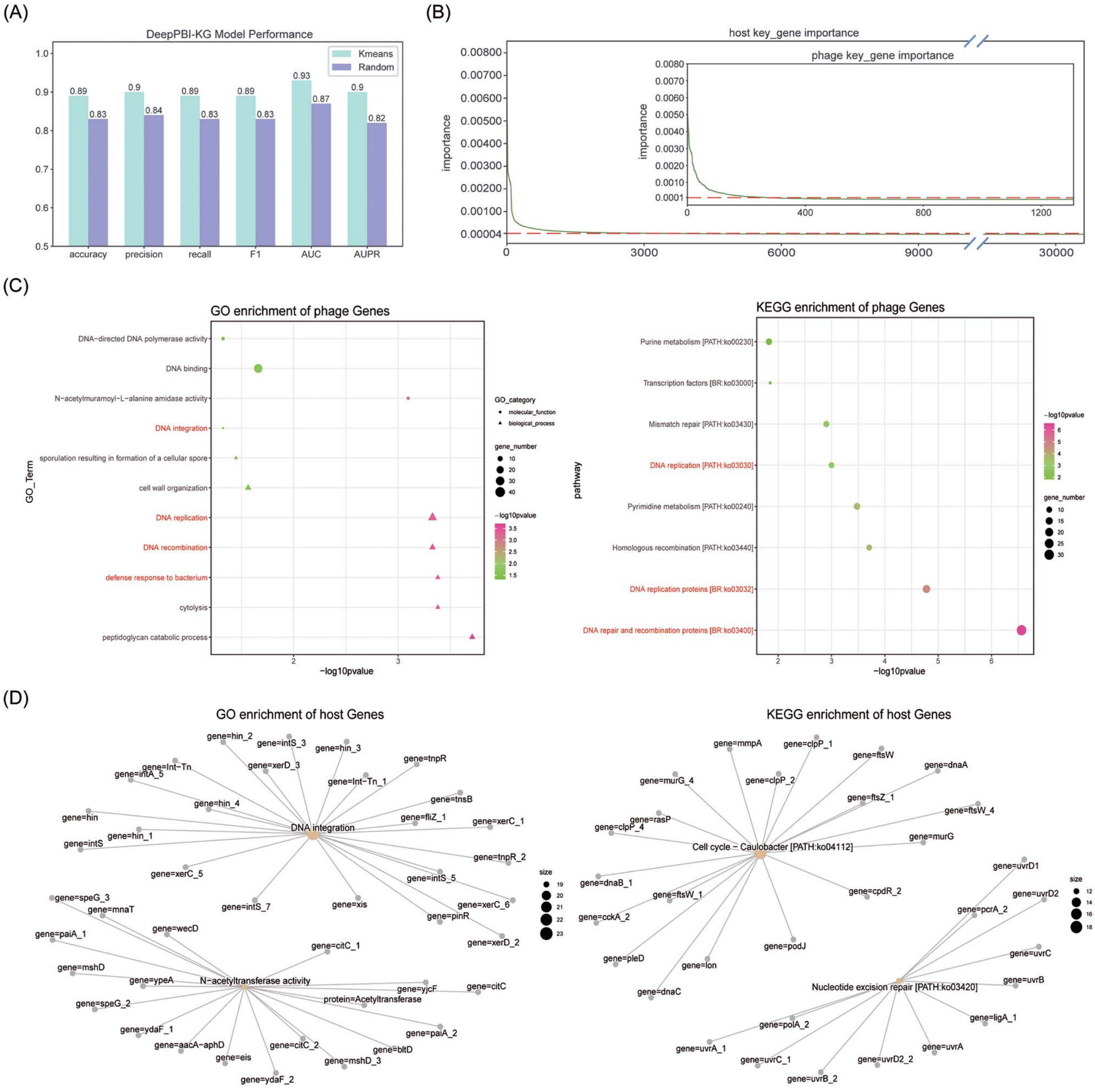


Figure 3. Model prediction performance, feature importance screening, and enrichment analysis results. (A) Performance of the model for each evaluation metric in the 2436 test sets. (B) Distribution of phage and bacterial feature importance and the division of phage and bacterial feature importance under the RF model. The red axis marks in the figure represent the contribution threshold set according to PCA. (C) GO enrichment analysis dot plot and KEGG enrichment analysis dot plot of phage genes with the top feature importance values. Important pathways are marked in red. (D) GO enrichment analysis network plot and KEGG enrichment analysis network plot of host genes with the top feature importance values. Important pathways are marked in red.

Table 1. Comparison of the performance of DeepPBI-KG and the other dataset. Comparison of the performance of DeepPBI-KG and DeepPBI-KG. Comparison of the performance of DeepPBI-KG and DeepPBI-RBP

Data	Accuracy	Precision	Recall	F1	AUC	AUPR
DeepPBI-KG	0.89	0.90	0.89	0.89	0.93	0.90
DeepPBI-WGS	0.84	0.87	0.84	0.84	0.93	0.88
DeepPBI-RBP	0.74	0.80	0.73	0.72	0.91	0.88

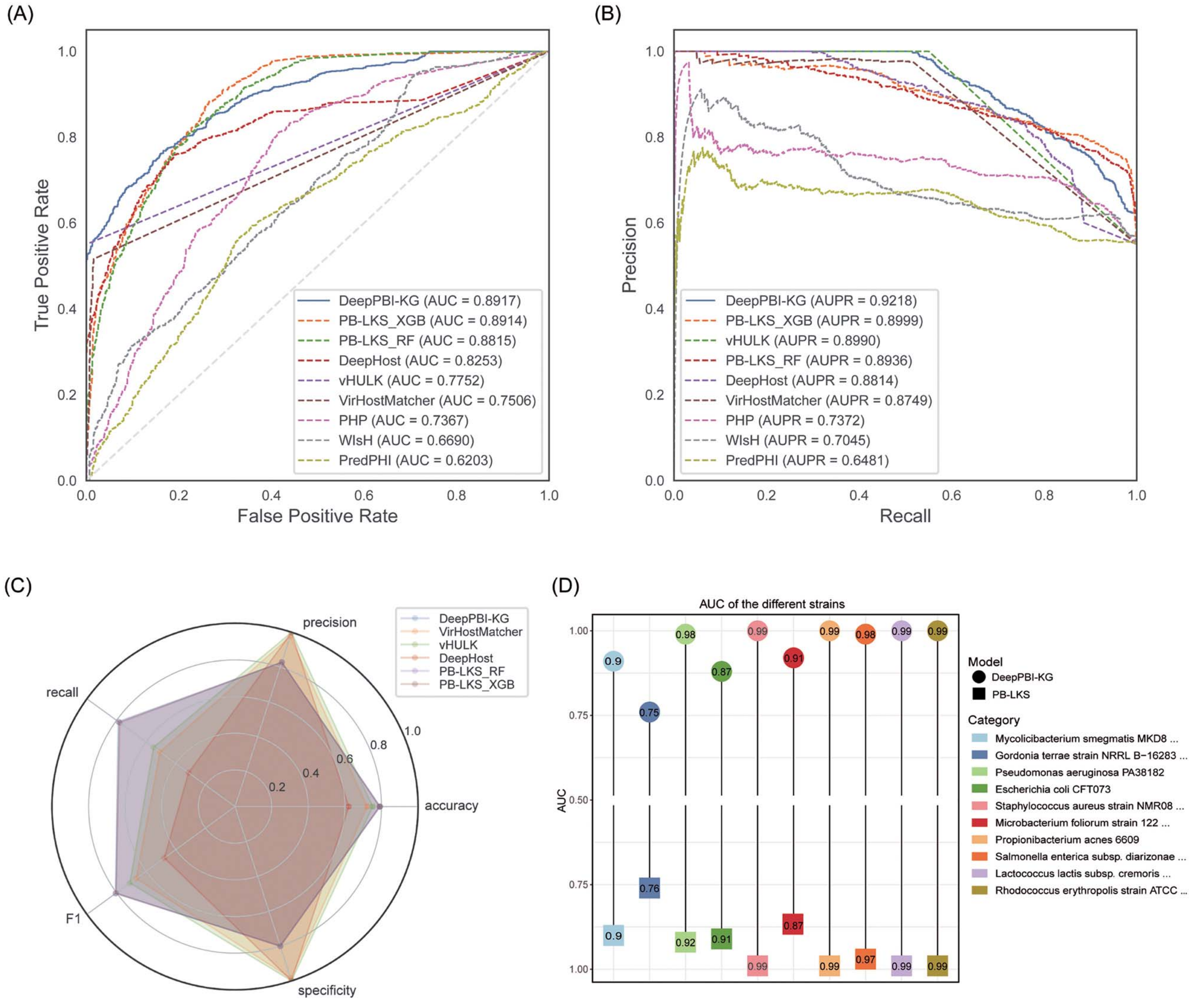


Figure 4. Comparison of DeepPBI-KG with state-of-the-art methods and AUC of each strain. (A) Comparison of the receiver operating characteristic (ROC) curve performance of different classifiers. (B) Comparison of the PR curves of different classifiers. (C) Radar map of different classifiers for accuracy, precision, recall, F1 score, and specificity. (D) The AUCs of DeepPBI-KG and PB-LKS were plotted for each strain. The circle under each bar represents DeepPBI-KG and the square represents PB-LKS.

considering various metrics, DeepPBI-KG outperformed all the other methods. The confusion matrix for all methods is shown in Supplementary Fig. S3.

AUC of each strain

To further verify the predictive efficacy of the model, the trained model was applied to predict interactions across various bacterial strains, with observation and evaluation of the AUC for each strain. Comparison with the SOTA methods showed that the performance of PB-LKS was nearly equivalent to that of DeepPBI-KG. The findings for PB-LKS, a tool for predicting phage-bacteria interactions, were published in January of this year in the journal ‘Briefings in Bioinformatic’. Hence, we also employed PB-LKS to predict the AUC for various strains and compared these results with those of our method. The AUC lollipop chart for the different strains revealed that half achieved an AUC above 0.9, with some even reaching 0.99 (Fig. 4D). The AUC effect was less pronounced but largely predictable due to insufficient data for the remaining strains. DeepPBI-KG and PB-LKS had consistent predictive outcomes for five strains and exhibited mixed performance for four strains, with the former outperforming the latter in one strain (*Pseudomonas aeruginosa* PA38182). Overall, the use of machine learning methods that leverage feature augmentation and data amplification based on DNA-protein characteristics helps enhance the ability to predict phage-bacteria interactions, enabling the prediction of interactions between phages and each host at a considerable AUC level.

Selection of DeepPBI-KG versus DeepPBI-WGS

It is conceivable that the predictive effect of DeepPBI-KG model depends on how many key genes the host has, so the genome integrity of the host should be considered. We plotted the host genome integrity distribution for the test set and divided genome integrity into three gradients according to the number of key genes: complete (800+), medium complete (400–800), and incomplete (1–400) (Supplementary Fig. S4A). The predicted results for each gradient under DeepPBI-KG and DeepPBI-WGS are shown in Supplementary Table S3. The accuracy of prediction decreased with decreasing gene integrity, but the overall prediction level was

Table 2. Comparison of the prediction effect of the retraining model based on the KP strain

Methods	Accuracy	Precision	Recall	F1	AUC	AUPR
DeepPBI-KG	0.9350	0.9268	0.9233	0.9250	0.9680	0.8879
DeepPBI-WGS	0.9350	0.9302	0.9191	0.9244	0.9368	0.8723
PB-LKS (RF)	0.8000	0.7264	0.6016	0.6581	0.8576	0.7937
PB-LKS (XGB)	0.8050	0.6812	0.7344	0.7068	0.8668	0.8073
PredPBI	0.8875	0.7862	0.8906	0.8352	0.9612	0.8735
PHP	0.8725	0.7770	0.8438	0.8089	0.8765	0.8993

still considerable, the accuracy was above 0.8, and the results for key genes were indeed better than those for the whole genome. Furthermore, we examined the prediction accuracy for hosts whose genome integrity was <200 to explore when key gene outcomes would be invalidated relative to genome-wide outcomes (Supplementary Fig. S4B). When the host genome integrity was <100, the prediction results for key genes gradually became worse than those for the whole genome. Therefore, currently, there is a higher degree of confidence in the prediction results obtained by selecting the whole genome.

Predictive potential of intraspecific interactions

We conducted a study on intraspecies interaction prediction using data from various pneumococcal strains detected by our research group during the treatment of hospital lung infections to assess the potential of DeepPBI-KG for intraspecies prediction. The pneumococcal host strains, although different, were all KP strains [31–33]. Some of the data for the interactions between phages with *KP* as the primary host and KP were confirmed through wet-laboratory experiments (for detailed information on the wet experiments, refer to the **Sequencing information and experimental validation of bacteriophage - *K. pneumoniae* interactions** section under **MATERIALS AND METHODS**). After preliminary data screening, phages and bacteria with overly short, incomplete, or inaccurate genomic sequences were eliminated. A total of 12 430 pairs of interacting samples from 110 KP phages and 113 KP subspecies were identified, including 3169 pairs of positive samples and 9261 pairs of samples with no interaction. The 110 × 113 interaction spectra of the KP strain were visualized as heatmaps and clustered (Supplementary Fig. S5). Most existing tools focus primarily on interspecies predictions, providing host predictions at the species and genus level without delving into specific subspecies. Only four tools, namely, PHP [10], PredPHI [6], WIsH [8], and PB-LKS [12], can predict intraspecies interactions, but only the first two tools allow the model to be custom retrained with a custom dataset. Consequently, we retrained the PHP and PredPHI models using KP data to predict interactions with KP strains and compared the training outcomes with those of DeepPBI-KG. Although PB-LKS, a newly released tool aimed at intraspecies prediction of phage-bacteria interactions, does not offer a custom training module, we utilized the source code and applied the PB-LKS concept for retraining with KP data, making predictions on KP data. After sample screening from the KP dataset, the training set comprised 2028 samples, with the remaining 400 samples extracted as an independent validation set (not present in the training set). The training set was divided into 7:3 ratios, with the initial 70% used for model training and the subsequent 30% used for model testing and evaluation. Our retraining results clearly surpassed those of all the other tools (Table 2). We trained models separately on the basis of the WGS (DeepPBI-WGS) and selected key genes (DeepPBI-KG) to further illustrate the importance of key genes. The performance of DeepPBI-KG also exceeded that of DeepPBI-WGS, reaffirming that key proteins can predict KP interactions with greater accuracy. Moreover, it was observed that the retrained PB-LKS performed notably poorly in predicting KP interaction, performing worse than PHP and PredPHI. Specifically, both the RF and XGB methods had the lowest AUC and AUPR values.

Conclusion

To further validate the crucial role of key genes in predicting interactions, genes contained in the RBP dataset were extracted. Following the prediction of feature importance via the RF model, the ordering of these genes within the sorted gene set was observed to determine whether the model predicted genes and proteins that truly influence interactions. Simultaneously, genes associated with significant GO/KEGG enrichment terms for both phages and hosts were extracted as described in the **Validation based on key genes** section. These genes, along with the gene set sorted after RF prediction of feature importance, were subjected to gene set enrichment analysis (GSEA). The aim was to observe the ordering of genes significantly enriched in the sorted gene set according to GO and KEGG enrichment. GSEA of phage genes indicated that within the queried gene set under GO/KEGG, only one pathway exhibited a significant p value, but each pathway tended to be upregulated. Similarly, the genes included in RBP dataset were not significantly differentially expressed but also tended to be upregulated (Supplementary Fig. S6A). The GSEA results of the host genes largely mirrored those of the phage genes, with host genes in the RBP_data not showing significance but exhibiting an increasing trend (Supplementary Fig. S6B). However, the P-value for the gene sets under the host's GO categories was all lower than those for the RBP dataset. The GSEA results for both phage and host genes suggest that the genes within RBP dataset indeed impact interactions. Nonetheless, the key genes screened by the RF model appear to be more reasonable and closer to the key genes that influence interactions. This is similar to using model fitting to approximate the theoretically optimal set of true interaction-influencing genes, with the model's 'fit' outperforming the RBP dataset. The GO/KEGG enrichment analysis of genes contained in the RBP dataset indicated that phage genes were primarily enriched in cell wall organization and peptidases and inhibitors (Supplementary Fig. S6C). The host genes were mainly enriched in siderophore uptake transmembrane transporter activity and signaling receptor activity (Supplementary Fig. S6D). Interestingly, Herridge *et al.* [30] reported in their 2020 review on the infection mechanism of *KP* and *Pneumococcus pneumoniae* that siderophores and lipopolysaccharides are key factors affecting host phage recognition. These key factors are involved in the recognition of host cell membrane surface receptors. These results seem to correlate with processes related to phage recognition of bacterial surface receptors. Moreover, the genes selected by the RF model primarily focus on functions such as DNA integration, replication, and phage invasion of host cells, which are related to the phage recognizing the host and infecting the host cell to inject its genetic material into the host cell. In summary, not all proteins that influence interactions are RBPs, and not all RBPs are crucial in mediating these interactions.

The increase in the incidence of bacterial infections has necessitated the emergence of various treatment methods, such as antibiotics. However, the misuse of antibiotics has led to increasing resistance among bacteria, making the use of phages a promising alternative treatment for bacterial infections [34–36]. Nonetheless, the rapid and accurate screening of phages capable of killing bacteria is time-consuming and labour-intensive, prompting the development of computational tools to predict phage-bacteria interactions [37]. However, existing predictive tools have limitations. Therefore, in this study, we propose DeepPBI-KG, a new method for predicting phage-bacteria interactions from a novel perspective, providing biological interpretability. Initially, this study leveraged augmented DNA-protein features rather than the whole genome to enrich a priori feature information from key genes. The benchmark results indicate that the DNA-protein features of key genes significantly contribute to predictive accuracy. Second, K-means negative sampling is employed to select high-quality negative samples, demonstrating that K-means sampling enhances the model's predictive performance. Finally, forward and reverse validation modules based on key_gene confirmed that the selected key genes are involved in phage-bacteria interactions. Enrichment analysis revealed that some genes were enriched in functions and pathways related to interactions, although the results from the negative validation module were not significant. The discrepancy arose because the RBP dataset was curated based on relevant literature and biological significance, potentially including genes that do not genuinely affect interactions. This also explains why there was an overall trend towards upregulation despite the nonsignificant GSEA results, indicating a discrepancy between the predicted and manually selected genes. Setting aside this error, genes predicted based on machine learning are still reliable to some extent.

Moreover, utilizing whole-genome predictions often involves considerable redundant information, which does not aid in further identification of which proteins and genes genuinely impact interactions. However, we can significantly improve prediction accuracy and precision by identifying key genes and critical proteins of phages and their hosts. Further interaction analysis, pinpointing the key genes and proteins responsible for killing bacteria, and then selecting specific proteins and genes for particular bacteria can help in the construction of super-phages. In response to antibiotics, many bacteria have evolved subspecies with enhanced multidrug resistance [38, 39]; thus, accurately and rapidly screening phages that kill subspecies will be more practical. Unfortunately, current predictive tools can predict only interspecies bacterial interactions and are not applicable for intraspecies predictions. Therefore, developing predictive models for intraspecies interactions is urgently needed. However, the experimental results of the phage-bacteria interaction predictions of the *KP* subspecies in this study suggest that DeepPBI-KG has potential for intraspecies predictions, with superior predictive performance compared to that of PB-LKS. In future work, we aim to continue increasing the dataset size to enhance the model prediction capabilities and further investigate intraspecies interaction patterns. This study, despite providing new insights into the prediction of phage-bacteria interactions, had certain limitations. First, the manually curated RBP dataset may not be accurate, resulting in nonsignificant enrichment results in the key_gene reverse validation module. The enrichment may be better than the results of this study. More accurate selection of phage tail proteins and host receptor proteins through extensive literature could yield enrichment results that more strongly support the hypothesis regarding key protein interactions. Second, the structural features applied in this study are tabular, so unstructured features such as protein sequence embedding should be considered in future research designs, and subsequent modelling and discussion should be attempted to address the predictive bottlenecks and limitations of existing methods.

Key Points

- A model is constructed on the basis of key genes, and a biological interpretation of the model is provided. Validation on the basis of key genes reveals that RBPs are not the only factors that influence this interaction.
- Intraspecific prediction is discussed. DeepPBI-KG exhibited good potential for predicting intraspecific interactions.
- The DeepPBI-KG prediction framework can be applied to any intraspecies prediction study, as long as sufficient intraspecies sequencing data and interaction information are provided.
- The Python package for DeepPBI-KG is freely available on GitHub (https://github.com/Tongqing-Wei/DeepPBI-KG).

Supplementary data

Supplementary data are available at *Briefings in Bioinformatics* online.

Conflict of interest: None declared.

Funding

This work was supported by grants from the National Key Research and Development Program of China (2021YFC2501800, 2022YFA0806200).

Data availability

All the sequencing data have been deposited in NGDC under submission number Batch0088600 in GWH and GB0003811 in GenBase (whole genome sequencing), BioProject accession number PRJCA025410. The data and scripts used are saved in GitHub https://github.com/Tongqing-Wei/DeepPBI-KG_data_process and https://github.com/Tongqing-Wei/DeepPBI-KG_external_data.

Author contributions

Tongqing Wei, Chenqi Lu, Wen-Hong Zhang, Xu Tao, and Ning Jiang performed the experiments, construct the model, analyzed the data, and wrote the manuscript. Xin Qi and Yankun Liu performed the sequencing and interaction experiments. Yi Zhang and Chen Chen provided the KP and phages strains. Hanxiao Du, Qianru Yang, Yutong Li and Yuanhao Tang helped to analyze data and revised the manuscript. Ning Jiang, Xu Tao and Wen-Hong Zhang supervised this project and are the corresponding authors. All authors have read the final manuscript and approved it for publication.

References

1. Khan S, Zakariah M, Rolfo C. *et al.* Prediction of mycoplasma hominis proteins targeting in mitochondria and cytoplasm of host cells and their implication in prostate cancer etiology. *Oncotarget* 2017;**8**:30830.
2. Dreyfuss D, Ricard J-D. Acute lung injury and bacterial infection. *Clin Chest Med* 2005;**26**:105–12.
3. Zhang C, Liu H, Sun L. *et al.* An overview of host-derived molecules that interact with gut microbiota. *iMeta* 2023;**2**:e88.
4. Toke O. Antimicrobial peptides: new candidates in the fight against bacterial infections. *Pept Sci Orig Res Biomol* 2005;**80**: 717–35.
5. 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.
6. 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.
7. Ruohan W, Xianglilan Z, Jianping W. *et al.* DeepHost: phage host prediction with convolutional neural network. *Brief Bioinform* 2022;**23**:bbab385.
8. 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.
9. Ahlgren NA, Ren J, Lu YY. *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.
10. Lu C, 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.
11. Amgarten D, Iha BKV, Piroupo CM. *et al.* vHULK, a new tool for bacteriophage host prediction based on annotated genomic features and deep neural networks. *Phage (New Rochelle)* 2020;**3**: 204–12.
12. Qiu J, Nie W, Ding H. *et al.* PB-LKS: A python package for predicting phage–bacteria interaction through local K-mer strategy. *Brief Bioinform* 2024;**25**:bbae010.
13. Zhou Z, Martin C, Kosmopoulos JC. *et al.* ViWrap: a modular pipeline to identify, bin, classify, and predict viral–host relationships for viruses from metagenomes. *iMeta* 2023;**2**: e118.
14. Altamirano FLG, Barr JJ. Unlocking the next generation of phage therapy: The key is in the receptors. *Curr Opin Biotechnol* 2021;**68**: 115–23.
15. Häuser R, Blasche S, Dokland T. *et al.* Bacteriophage protein–protein interactions. *Adv Virus Res* 2012;**83**:219–98. https://doi.org/10.1016/B978-0-12-394438-2.00006-2.
16. Boeckaerts D, Stock M, Criel B. *et al.* Predicting bacteriophage hosts based on sequences of annotated receptor-binding proteins. *Sci Rep* 2021;**11**:1467.
17. Pruitt KD, Tatusova T, Maglott DR. NCBI reference sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res* 2005;**33**: D501–4.
18. Gao NL, Zhang C, Zhang Z. *et al.* MVP: a microbe–phage interaction database. *Nucleic Acids Res* 2018;**46**:D700–7. https://doi.org/10.1093/nar/gkx1124.
19. 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.
20. Seemann T. Prokka: Rapid prokaryotic genome annotation. *Bioinformatics* 2014;**30**:2068–9. https://doi.org/10.1093/bioinformatics/btu153.
21. Bertozzi Silva J, Storms Z, Sauvageau D. Host receptors for bacteriophage adsorption. *FEMS Microbiol Lett* 2016;**363**:fnw002.
22. Wang B, Mei C, Wang Y. *et al.* Imbalance data processing strategy for protein interaction sites prediction. *IEEE/ACM Trans Comput Biol Bioinform* 2019;**18**:985–94.
23. Genome warehouse: a public repository housing genome-scale data. *Genomics Proteomics Bioinformatics* 2021;**19**:584–9. https://doi.org/10.1016/j.gpb.2021.04.001.
24. Database resources of the national genomics data center, China National Center for bioinformation in 2024. *Nucleic Acids Res* 2024;**52**:D18–32. https://doi.org/10.1093/nar/gkad1078.
25. Andrew MK, Amanda M, Thomas EW. *et al.* Eumeration of bacteriophages by double agar overlay plaque assay. *Methods Mol Biol* 2009;**501**:69–76.
26. Camarillo-Guerrero LF, Almeida A, Rangel-Pineros G. *et al.* Massive expansion of human gut bacteriophage diversity. *Cell* 2021;**184**:e1099.
27. Chibani-Chennoufi S, Bruttin A, Dillmann M-L. *et al.* Phage-host interaction: an ecological perspective. *J Bacteriol* 2004;**186**: 3677–86.
28. Chen Z, Zhao P, Li F. *et al.* iFeature: a python package and web server for features extraction and selection from protein and peptide sequences. *Bioinformatics* 2018;**34**:2499–502. https://doi.org/10.1093/bioinformatics/bty140.
29. Labrie SJ, Samson JE, Moineau S. Bacteriophage resistance mechanisms. *Nat Rev Microbiol* 2010;**8**:317–27. https://doi.org/10.1038/nrmicro2315.
30. Herridge WP, Shibu P, O'Shea J. *et al.* Bacteriophages of *Klebsiella spp.*, their diversity and potential therapeutic uses. *J Med Microbiol* 2020;**69**:176.
31. Cao F, Wang X, Wang L. *et al.* Evaluation of the efficacy of a bacteriophage in the treatment of pneumonia induced by multidrug resistance *Klebsiella pneumoniae* in mice. *Biomed Res Int* 2015;**2015**:752930.
32. Kumari S, Harjai K, Chhibber S. Isolation and characterization of *Klebsiella pneumoniae* specific bacteriophages from sewage samples. *Folia Microbiol* 2010;**55**:221–7. https://doi.org/10.1007/s12223-010-0032-7.
33. Tabassum R, Shafique M, Khawaja KA. *et al.* Complete genome analysis of a *Siphoviridae phage TSK1* showing biofilm removal potential against *Klebsiella pneumoniae*. *Sci Rep* 2018;**8**:17904.
34. Chadha P, Katare OP, Chhibber S. In vivo efficacy of single phage versus phage cocktail in resolving burn wound infection in BALB/c mice. *Microb Pathog* 2016;**99**:68–77. https://doi.org/10.1016/j.micpath.2016.08.001.
35. Gu J, Liu X, Li Y. *et al.* A method for generation phage cocktail with great therapeutic potential. *PLoS One* 2012;**7**:e31698.
36. Strathdee SA, Hatfull GF, Mutalik VK. *et al.* Phage therapy: from biological mechanisms to future directions. *Cell* 2023;**186**:17–31. https://doi.org/10.1016/j.cell.2022.11.017.
37. 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.
38. Nikaido H. Multidrug resistance in bacteria. *Annu Rev Biochem* 2009;**78**:119–46. https://doi.org/10.1146/annurev.biochem.78.082907.145923.
39. Van Duin D, Paterson DL. Multidrug-resistant bacteria in the community: trends and lessons learned. *Infect Dis Clin* 2016;**30**: 377–90. https://doi.org/10.1016/j.idc.2016.02.004.