Briefings in Bioinformatics, 00(0), 2021, 1–10

https://doi.org/10.1093/bib/bbab348
Problem Solving Protocol

PHIAF: prediction of phage-host interactions with GAN-based data augmentation and sequence-based feature fusion

Menglu Li and Wen Zhang

Corresponding author: Wen Zhang, College of Informatics, Huazhong Agricultural University, Wuhan 430070, China. E-mail: zhangwen@mail.hzau.edu.cn

Abstract

Phage therapy has become one of the most promising alternatives to antibiotics in the treatment of bacterial diseases, and identifying phage-host interactions (PHIs) helps to understand the possible mechanism through which a phage infects bacteria to guide the development of phage therapy. Compared with wet experiments, computational methods of identifying PHIs can reduce costs and save time and are more effective and economic. In this paper, we propose a PHI prediction method with a generative adversarial network (GAN)-based data augmentation and sequence-based feature fusion (PHIAF). First, PHIAF applies a GAN-based data augmentation module, which generates pseudo PHIs to alleviate the data scarcity. Second, PHIAF fuses the features originated from DNA and protein sequences for better performance. Third, PHIAF utilizes an attention mechanism to consider different contributions of DNA/protein sequence-derived features, which also provides interpretability of the prediction model. In computational experiments, PHIAF outperforms other state-of-the-art PHI prediction methods when evaluated via 5-fold cross-validation (AUC and AUPR are 0.88 and 0.86, respectively). An ablation study shows that data augmentation, feature fusion and an attention mechanism are all beneficial to improve the prediction performance of PHIAF. Additionally, four new PHIs with the highest PHIAF score in the case study were verified by recent literature. In conclusion, PHIAF is a promising tool to accelerate the exploration of phage therapy.

Key words: phage-host interactions; generative adversarial network; feature fusion; attention mechanism.

Introduction

Available reports have demonstrated the possible involvement of bacterial infection in the growth and development of various types of diseases, including cholera [1], inflammatory bowel disease [2], colon cancer [3–5], tetanus [6] and different types of cancer [7]. Researchers discovered antibiotics in 1928 and have since used them in clinical practice to treat serious bacterial diseases and save countless lives [8]. Unfortunately, due to the overuse of antibiotics, bacteria have developed a few resistance mechanisms [9]. In 2019, the US Centers for Disease Control and Prevention reported that approximately 2.8 million cases of antibiotic-resistant infections occur each year in the United States, resulting in more than 35 000 deaths [10]; in Europe, about 33 000 people die from antibiotic-resistant infections each year [11]. Thus, it is urgent to develop new antibiotics or alternative therapies to avoid further deterioration of antibiotic-resistant infections. However, many pharmaceutical companies no longer develop new antibiotics because of their high production costs, unsatisfactory expected benefits and long research and development time [12, 13]. Therefore, researchers want to look for alternative therapies to reduce antibiotic-resistant infections and treat bacterial diseases.

Bacteriophages can not only destroy specific bacteria hosts but also replicate exponentially, and these characteristics make bacteriophages one of the most promising therapies in the treatment of bacterial diseases and address antibiotic-resistant infections [14]. Determining phage-host interactions (PHIs) helps to

Menglu Li is a PhD candidate in the College of Informatics at Huazhong Agricultural University.
Wen Zhang is a professor in the College of Informatics at Huazhong Agricultural University.

Submitted: 2 January 2021; **Received (in revised form):** 5 July 2021

understand whether phages can be used to treat bacterial diseases. However, experimental verification of PHIs requires considerable time, manpower and money. Therefore, researchers have attempted to develop computational PHI prediction methods to screen out target phages for treating bacterial diseases and to guide the *in vivo* validation, thereby greatly reducing the required time and costs [15].

Molecular and ecological coevolutionary processes shape phage and bacterial genomes and leave signals in their genomic sequences that allow researchers to predict PHIs [15], so various PHI computational methods based on phage and host genomic sequences have been developed [16–18]. For example, Ahlgren *et al.* [19] proposed VirHostMatcher (VHM) based on DNA sequences to predict PHIs by calculating the distance between the oligonucleotide frequency patterns of phages and hosts. However, the running time of VHM hindered its development on large datasets, so Galiez *et al.* [20] proposed WIsH to reduce the running time by constructing a Markov model to predict the prokaryotic host of bacteriophages. Compared with that of VHM, the running time of WIsH was reduced by a factor of several hundred. In addition to VHM and WIsH, which predict PHIs by calculating the similarity between phages and hosts, researchers have used various machine learning classifiers, including logistic regression (LR), support vector machine (SVM), random forest (RF) and naive Bayesian (NB), to predict PHIs [21]. Further, PHP [22] and VIDHOP [23] were developed to enhance the PHI prediction performance. PHP trained a Gaussian model by calculating the differences in k-mer frequencies between viral and host genomic sequences, and VIDHOP used deep neural networks to predict phages related to three different viruses (influenza A virus, rabies lyssavirus and rotavirus A).

Some studies have shown that proteins play a fundamental role in the biological processes of phages and hosts [24, 25]; thus, researchers have proposed PHI prediction methods based on protein sequences [26, 27]. For example, Leite *et al.* [28, 29] utilized the primary structure sequences from phage and host proteins and classic classifiers, including RF, SVM, LR, k-nearest neighbor (KNN), multi-layer perceptron (MLP) and NB, to predict PHIs. On the basis of the above method, Li *et al.* [30] used a convolutional neural network (CNN) to improve the performance of PHI prediction.

Although existing methods achieve good performance in PHI prediction, some challenges remain. First, there are thousands of experimentally verified PHIs in databases [31–34], but only a few hundred non-redundant PHIs are available and can be used to build predictive models. This limitation hinders the development of predictive models with high performance. Second, most existing methods use either the DNA sequences or protein sequences of phages and hosts to construct predictive models but rarely combine two types of sequences. Third, although a variety of features and machine learning techniques have been used to build prediction models, these models often lack sufficient interpretability, which obstructs elaborating the mechanism of PHIs.

In recent years, deep learning technology has received extensive attention in the field of bioinformatics, and researchers have applied such techniques to handle different tasks [35, 36]. The generative adversarial network (GAN), as a branch of deep learning technology, was originally used for image processing [37, 38] and later showed excellent performance in data augmentation. For instance, Wan *et al.* [39] successfully used a GAN to generate biophysical features based on protein sequences. Meanwhile, researchers developed an attention mechanism [40] for deep learning to increase the interpretability of predictive models and to improve prediction performance. The development of these deep learning technologies motivates us to further enhance and improve PHI prediction.

In the current study, we propose a novel PHI prediction method, abbreviated as PHIAF, based on GAN data augmentation and sequence-based feature fusion to solve the various challenges of PHI prediction. First, PHIAF uses GAN to construct a data augmentation module, which generates high-quality pseudo samples to overcome the bottleneck of the PHI data scarcity. Second, PHIAF fuses different features encoded by the DNA and protein sequences of phages and hosts to enhance the prediction performance. Third, PHIAF utilizes CNN to build a PHI prediction module and incorporates an attention mechanism into CNN to provide interpretability of the prediction model. Experimental results show that PHIAF is superior to the state-of-the-art methods of PHI prediction. The ablation study and discussion indicate that the pseudo samples generated by the data augmentation module, the fusion of DNA and protein sequence-derived features and the attention mechanism in CNN effectively improve the performance of PHIAF.

Algorithm 1 : Data processing to remove redundant phages.

Require: The set of hosts, $H = \{h_1, h_2, ..., h_n\}$, n is the number of hosts; the set of phages corresponding to different hosts, $P = \{P_1, P_2, ..., P_n\}$, where $P_i = [p_1, p_2, ..., p_m]$ is the set of phages of host h_i, $i \in [1, n]$; similarity matrix of different phages, $S = \{s_{p_1,p_2}, s_{p_1,p_3}, ..., s_{p_{m-1},p_m}\}$, m is the number of phages;
Ensure: The interactions between non-redundant phages and hosts, I;

```
function main($P, H, S$)
  $I \leftarrow []$
  for $k \leftarrow 1$ to $n$ do
    $I_k \leftarrow []$
    $I_k \leftarrow del\_redundant(P_k)$
    $I = I + I_k$
  end for
  return $I$
end function
function del_redundant($P_k$)
  $P_{del} \leftarrow [], P_{new} \leftarrow []$
  $I_k \leftarrow I_k + (P_k[0], h_k)$
  $P_{del} \leftarrow P_{del} + P_k[0]$
  for $j \leftarrow 1$ to $m$ do
    if $s_{P_k[0],P_k[j]} > 0.90$ then
      $P_{del} \leftarrow P_{del} + P_k[j]$
    end if
  end for
  $P_{new} \leftarrow P_k - P_{del}$
  if $length(P_{new}) > 1$ then
    $del\_redundant(P_{new})$
  end if
  if $length(P_{new}) = 1$ then
    $I_k \leftarrow I_k + (P_{new}[0], h_k)$
  end if
  return $I_k$
end function
```

Materials and methods

Dataset

We download data (including phages, hosts and their interactions) from four widely used databases on March 2021, including

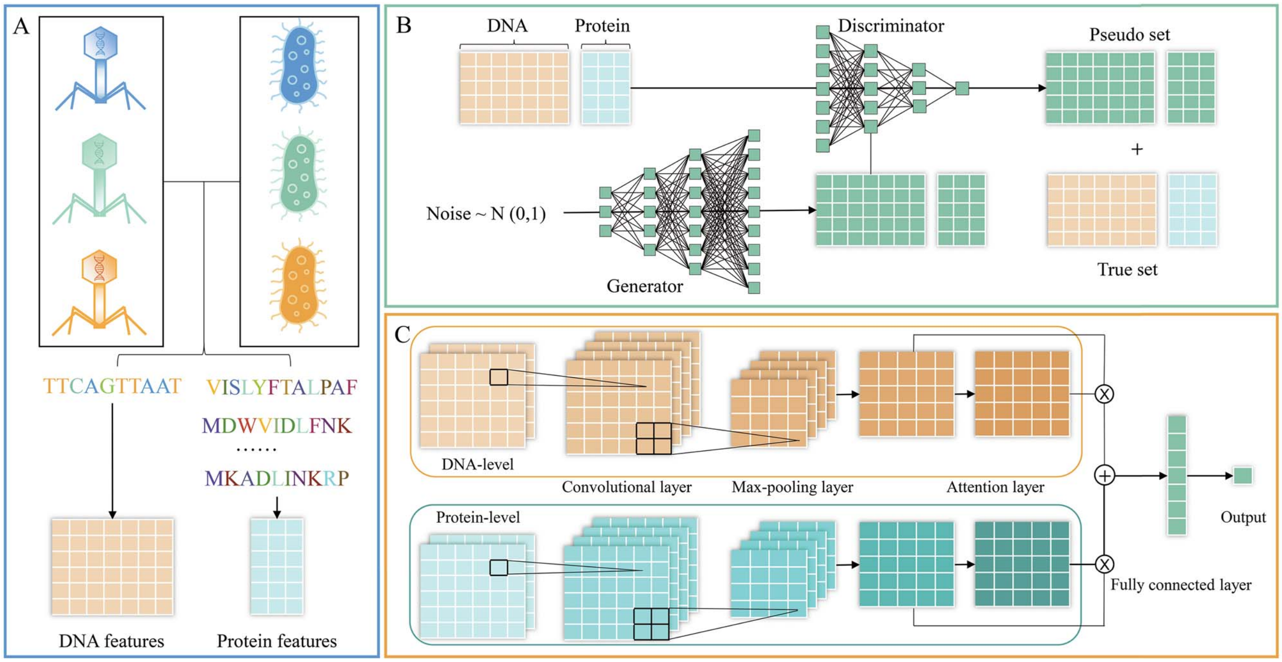

Figure 1. An overview of PHIAF. (**A**) Feature extraction module. (**B**) Data augmentation module. (**C**) PHI prediction module.

MVP [31], PhagesDB [32], VHDB [33] and NCBI [34], and merge these data to construct a dataset with more PHIs for our study. The data in these four databases are subjected to the following processing. First, we delete the PHIs that are not published in the literature or not included in NCBI records to ensure that the PHIs are reliable. Second, we remove data incorrectly marked as phages/hosts based on the definition of phages (phages are viruses that infect and replicate within bacteria and archaea). The removed data include phages that do not belong to viruses and hosts that do not belong to bacteria or archaea. Third, we extract whole-genome sequences and coding protein sequences from the NCBI database for remaining phages and hosts.

After the above processes, we combine the remaining phages and hosts of the four databases and remove duplicates, resulting in a total of 5399 interactions between 5331 phages and 235 hosts. The number of phages is much larger than the number of hosts, and one host may interact with multiple phages. We use Algorithm 1 to remove redundant phages with high similarity for each host (the comparison between different similarity measures and the impact of phage $P_k[0]$ on prediction performance are provided in Sections 1 and 2 of Supplementary Materials, respectively). We set 0.90 as a high similarity threshold, which is the same as the default threshold of the CD-HIT tool [41]. After redundancy reduction, we obtain a benchmark dataset with 312 interactions between 304 phages and 235 hosts, which can be used to better evaluate the performance of the prediction models. In this dataset, we set 312 known PHIs as positive samples and select negative samples from all unknown PHIs while ensuring that the numbers of positive and negative samples are equal.

PHIAF

PHIAF consists of three main modules: feature extraction, data augmentation and PHI prediction. A schematic diagram of PHIAF is shown in Figure 1. First, DNA and protein sequences of phages and hosts are encoded into features (Figure 1A). Second, a GAN-based data augmentation module is used to generate pseudo PHIs (Figure 1B). Finally, a PHI prediction module is built under a CNN framework with attention to utilize the features derived from DNA and protein sequences after reshaping into appropriate forms to predict PHIs (Figure 1C).

Feature extraction module

Some studies have shown that DNA and protein sequences play a fundamental role in the biological evolution of bacteriophages and hosts. To obtain more comprehensive and effective information, we extract features from DNA and protein sequences of phages and hosts, as summarized in Table 1; detailed information of these features is provided in Section 3 of Supplementary Materials.

Since DNA sequences of phages and hosts have different lengths (analysis of the distribution of the DNA and protein sequence lengths of phages and hosts is provided in Section 4 of Supplementary Materials), we consider several DNA sequence-derived features that are unrelated to sequence length, including Kmer, reverse compliment Kmer (RCKmer), nucleic acid composition (NAC), di-nucleotide composition (DNC), tri-nucleotide composition (TNC), the composition of k-spaced nucleic acid pairs (CKSNAP) and electron-ion interaction pseudopotentials of trinucleotide (PseEIIP). We calculate these features with the software iLearn [42] and obtain a 340-dimensional feature vector for each phage/host DNA sequence.

We follow previous studies [28–30] to extract widely used protein sequence-derived features, including amino acid composition (AAC), the abundance of chemical elements composing a protein (AC) and the molecular weight of a protein (MW). Since each phage/host has multiple protein sequences, we consider six operators (mean, maximum, minimum, standard deviation, variance and median) to integrate features from protein sequences and obtain a 162-dimensional feature vector for each phage/host.

Table 1. Description of the features that originated from DNA and protein sequences

Levels	Features	Descriptions
DNA	Kmer	the occurrence frequencies of k neighboring nucleic acids ($k = 3$)
	RCKmer	a variant of Kmer, which removes the reverse complement Kmers
	NAC	the frequency of each nucleic acid type (A, C, G, T) in a nucleotide sequence
	DNC	the frequency of two nucleic acid types in a nucleotide sequence
	TNC	the frequency of three nucleic acid types in a nucleotide sequence
	CKSNAP	the frequency of nucleic acid pairs separated by any p nucleic acid ($p = 5$)
	PseEIIP	mean EIIP values (A: 0.1260, C: 0.1340, G: 0.0806, T: 0.1335) of trinucleotides in each sequence
Protein	AAC	the frequency of each amino acid in a protein sequence
	AC	abundance of selected chemical elements composing a protein
	MW	molecular weight of a protein sequence

Data augmentation module

The GAN [37] is a new type of generative model that aims to generate high-quality pseudo samples by precisely learning the underlying distribution of real samples. This model has received considerable attention and achieved outstanding performance in many fields [43, 44]. In this study, we use a GAN [39] to address the data scarcity of PHIs in our dataset for model training.

We first set the real positive samples as $I = \{(p_1, h_1), (p_2, h_2), \ldots, (p_m, h_n)\}$, $V = \{V_{p_1,h_1}, V_{p_2,h_2}, \ldots, V_{p_m,h_n}\}$ to represent the feature vectors of these samples, which are composed of the DNA and protein features of phages and hosts encoded above. m and n represent the numbers of phages and hosts, respectively. These feature vectors of positive samples (V) are input into the GAN to generate high-quality pseudo feature vectors of samples, where the GAN is composed of two neural networks (generator and discriminator) that 'fight' against each other to learn the distribution of real samples. One network (generator) tries to generate pseudo samples via five fully connected layers (formula 1). Another network (discriminator), which is composed of four fully connected layers (formula 2), tries to distinguish whether a given sample is real. Each network's task gets better and better until equilibrium is reached, where the generator cannot make better samples, and the discriminator cannot separate real and pseudo samples.

$$O_{ge} = FC_t\left(FC_r\left(FC_r\left(FC_r\left(FC_r(V)\right)\right)\right)\right) \tag{1}$$

$$O_{di} = FC_l\left(FC_l\left(FC_l\left(FC_l(V', V)\right)\right)\right) \tag{2}$$

where V' is the feature vectors of pseudo samples, O_{ge} is the generator output, O_{di} is the discriminator output, FC_l (or FC_r) represents MLP with the LeakyReLU (or ReLU) activation function and FC_t means MLP with the Tanh activation function.

We conduct the classifier two-sample tests (C2ST) method [45] using the KNN ($k = 1$) and leave-one-out cross-validation (LOOCV) to distinguish the real and pseudo samples. The C2ST method involves accepting or rejecting a null hypothesis of P being equal to Q (where P and Q are the distributions of two equal-sized sets of samples). If the null hypothesis is accepted, the classification accuracy for predicting the binary labels of held-out samples will be near the level of chance (that is 0.50). Therefore, the real and pseudo samples are indistinguishable when the KNN classification accuracy is the closest to 0.50 under LOOCV. Finally, we choose the indistinguishable pseudo samples to amplify our dataset, represented by $I' = \{(p'_1, h'_1), (p'_2, h'_2), \ldots, (p'_m, h'_n)\}$.

The above processing is used to amplify positive samples in our dataset. Since all unknown PHIs are candidate negative samples and are much more than positive samples, we randomly select negative samples from this candidate set to ensure that the numbers of positive and negative samples are equal. Finally, we combine the real positive samples, selected negative samples and pseudo positive samples to construct an augmented dataset, which is used to train the prediction model.

PHI prediction module

Based on the augmented dataset, we use the DNA and protein sequence-derived features of phages and hosts to build the prediction model.

As described in the previous section, we have two types of feature vectors for each phage/host, which are extracted from DNA and protein sequences, respectively. Since sequence-derived features usually have complicated short-/long-range dependency, we reshape the DNA/protein feature vectors into 'images' to capture the complicated relationship between their dimensions [46]. We first adopt the Min-Max normalization to normalize values in feature vectors to the range from 0 to 1. Let N denote the dimension of the feature vector of a DNA/protein; then, we reshape the feature vector into an $n \times n$ 'image' by placing values by row, where n satisfies the condition: $(n-1) \times (n-1) < N$ and $N \le n \times n$. When $N < n \times n$, we implement padding by adding zeros to the remaining $n \times n - N$ entries. Finally, we obtain the DNA and protein sequence-derived feature matrices for each phage (or host), denoted as $\mathbf{M}_d^P$ and $\mathbf{M}_p^P$ ($\mathbf{M}_d^H$ and $\mathbf{M}_p^H$), respectively.

Then, we construct a bi-level architecture (DNA- and protein-level) to extract deeper features from the DNA and protein feature matrices. For the DNA-level, we stack the DNA-derived feature matrix of phages and hosts across channels to form a combined matrix and then input this combined matrix into a two-layer CNN to produce a feature map O_d with more meaningful information. The CNN includes a convolutional layer and a max-pooling layer.

$$O_d = MaxPool\left(Conv2D([\mathbf{M}_d^P, \mathbf{M}_d^H])\right) \tag{3}$$

where $Conv2D$ and $MaxPool$ represent the convolutional layer and max-pooling layer, respectively, and $[\cdot, \cdot]$ represents stacking across channels. Similarly, the protein-derived feature matrices of phages and hosts are combined across channels and fed to the same CNN to produce a feature map O_p at the protein-level.

$$O_p = MaxPool\left(Conv2D([\mathbf{M}_p^P, \mathbf{M}_p^H])\right) \tag{4}$$

The attention mechanism aims to imitate the action of the human brain to selectively concentrate on a few important parts while ignoring others in machine learning tasks [47]. The DNA- and protein-level features we have may make different contributions to the PHI prediction. Thus, we introduce an attention mechanism into our model, add an attention layer to capture important features and then integrate these features.

In the attention layer, the DNA- and protein-level feature maps (O_d and O_p) are input into a fully connected layer to calculate weight vectors (α_d and α_p), respectively; then, the feature maps are multiplied by the corresponding weight vectors. Finally, the output of the attention layer is calculated as follows:

$$\mathrm{O}_{att} = \mathrm{O}_d \otimes \alpha_d + \mathrm{O}_p \otimes \alpha_p \tag{5}$$

where ⊗ is the element-wise multiplication. At last, the output of attention layer is feed into a two-layer MLP to yield the probability of samples being a PHI.

$$Pred = FC_s\left(FC_r\left(\mathrm{O}_{att}\right)\right) \tag{6}$$

where FC_s means the MLP with Sigmoid activation function.

PHIAF optimization

There are several important hyper-parameters in the data augmentation and PHI prediction modules of PHIAF. In the data augmentation module, we set the numbers of neurons in the different fully connected layers of the generator as 128, 256, 512, 1024 and 1004 and those in the discriminator as 512, 256, 128 and 1, respectively. In the PHI prediction module, we set 32 filters and the size of a filter as 3 × 3 in the convolutional layer and set the size of the max-pooling layers of the DNA- and protein-level as 3 × 3 and 2 × 2, respectively. The difference in the size of these two max-pooling layers ensures that the dimensions of O_d and O_p are equal. We also consider the learning rate (optional values are 0.1, 0.01, 0.001 and 0.0001), batch size (optional values are 8, 16, 32, 64 and 128), loss rate of the dropout layer (optional values include 0.25, 0.5 and 0.75) and the number of neurons in the fully connected layer (optional values include 16, 32, 64, 128, 256 and 512) to determine the optimal parameters. The model performance given different parameters is provided in Section 5 of Supplementary Materials.

In addition, we utilize dropout and batch normalization layers in the PHI prediction module to prevent overfitting and improve generalizability [48, 49]. The data augmentation module and PHI prediction module are trained independently. Additionally, we adopt the Wasserstein loss add gradient penalty as the loss function for the data augmentation module and use the binary cross-entropy loss function for the PHI prediction module. All of the above loss functions are optimized by the Adam optimizer [50].

Results

Performance assessment

In this study, we adopt 5-fold cross-validation (5-CV) to evaluate the performance of PHIAF. All samples in our dataset are randomly divided into five equal-sized subsets. The cross-validation process is repeated five times, and every subset is used as the test set in turn while the remaining four subsets are used as the training set. The final 5-CV results are generated by averaging the five test set results. We employ several commonly used evaluation measures [51, 52] to assess the performance of PHIAF and state-of-the-art methods, including specificity (Spe), sensitivity (Sen), accuracy (Acc), F1-score (F1), area under the receiver-operating characteristic curve (AUC) and area under the precision-recall curve (AUPR). The measures are calculated as follows:

$$Spe = \frac{TN}{TN + FP} \tag{7}$$

$$Sen = \frac{TP}{TP + FN} \tag{8}$$

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \tag{9}$$

$$F1 = \frac{2 * TP}{2 * TP + FP + FN} \tag{10}$$

where *TP* (*TN*) denotes the number of positive (negative) samples correctly classified in the prediction and *FP* (*FN*) represents the number of incorrectly identified negative (positive) samples.

Comparison with state-of-the-art methods

To demonstrate the effectiveness of PHIAF, we compare it with the following state-of-the-art methods, including DNA sequence-based algorithms and protein sequence-based methods.

- VHM [19] is a DNA sequence-based method that computes the distance between the oligonucleotide frequency patterns of phages and hosts and obtains the possibility of PHIs based on this distance.
- WIsH [20] is a method based on DNA sequences that trains a Markov model for each host and calculates the probability of interactions for all phages.
- PHP [22] is a DNA sequence-based method that constructs a Gaussian model to predict PHIs using k-mer frequencies between virus and host genomic sequences.
- RF, SVM, KNN and MLP are widely used machine learning classifiers; Leite *et al.* [28] utilizes protein sequence-derived features and these classifiers to predict PHIs.
- PredPHI [30] is a protein sequence-based method that predicts PHIs under a CNN framework.

We first compare the PHIAF with these methods using 5-CV. As shown in Figure 2, PHIAF outperforms all comparison methods in terms of AUC and AUPR. In general, the DNA sequence-based methods (VHM, WIsH and PHP) produce better results than the protein sequence-based methods (RF, SVM, KNN, MLP and PredPHI), indicating that the information from the DNA sequences may play a more important role in PHIs. Our PHIAF model, which fuses information originating from DNA and protein sequences, is superior to models based on either DNA sequences or protein sequences, achieving 13.63% and 14.75% improvement, on average, in terms of the AUC and AUPR.

The predictive capability of models for unseen data is also important. Thus, we randomly select one-third of the phages and hosts in our dataset and use them and their interactions as a test set. The remaining phages and hosts and their interactions are used to train the prediction models. Under this experimental setting, phages or hosts in the test set are not included in

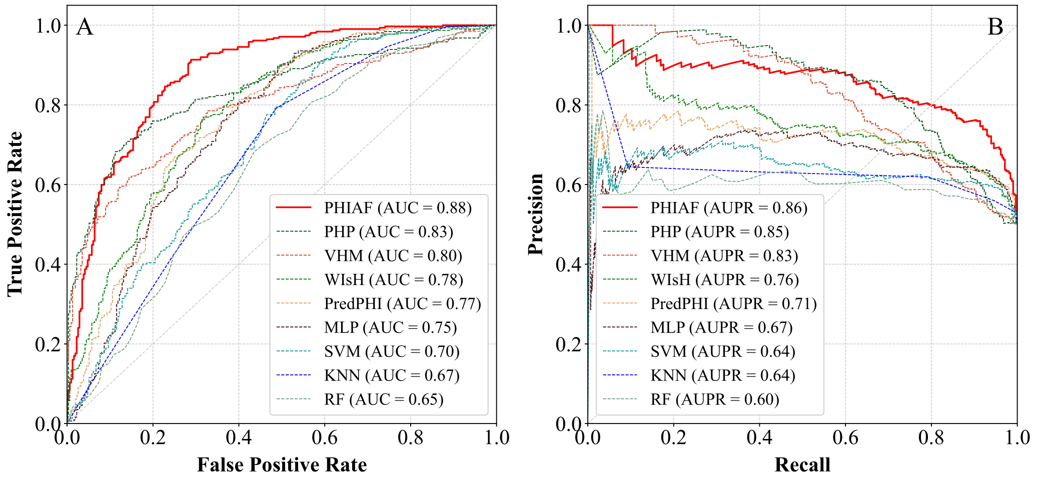

Figure 2. The performance of PHIAF and state-of-the-art methods using 5-fold cross-validation.

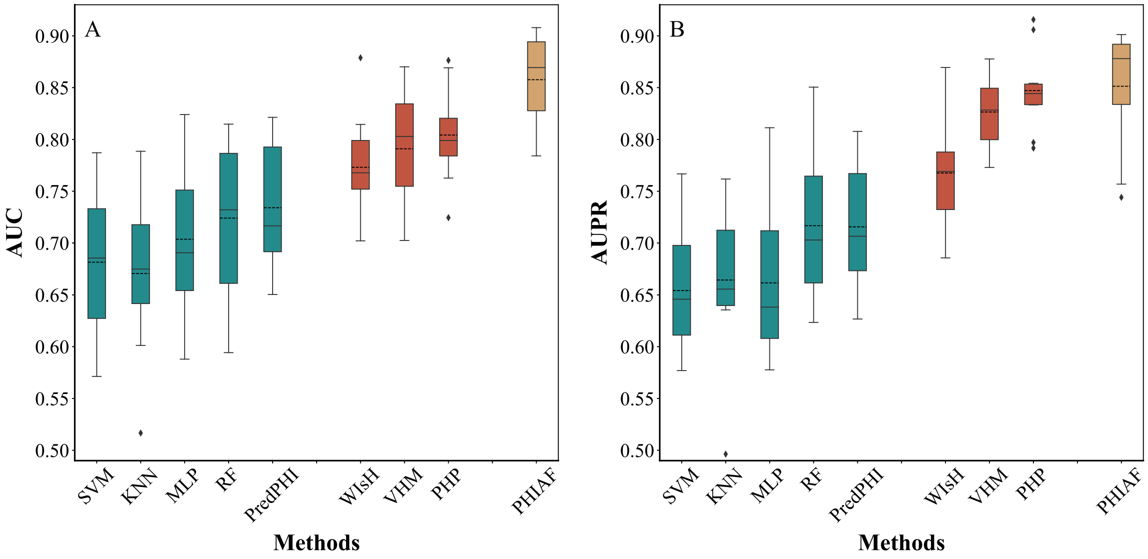

Figure 3. The performance of PHIAF and state-of-the-art methods for the unseen data. The solid line is the median of results, and the dashed line represents the mean.

the training set, and they are taken as unseen data. We train the prediction model based on the training set and then apply the trained model to the test set to evaluate the performance of the prediction model on unseen data. To avoid evaluation bias, we repeat the training and testing 10 times and adopt the average/median performance. As shown in Figure 3, PHIAF outperforms the other methods, generates higher average AUC (0.86) and AUPR scores (0.85) as well as reasonable standard deviation, and the medians of the AUC and AUPR scores are 0.88 and 0.87. Similar to the 5-CV results, DNA sequence-based methods produce better results than protein sequence-based methods. Moreover, a comparison between the results of the 5-CV and testing on unseen data indicates that our method achieves very similar performance in both cases (the AUC and AUPR differ by 2% and 1%), and these results demonstrate that our proposed method is robust and can perform well on unseen data, indicating that PHIAF is a promising tool for identifying PHIs from sequence data.

Ablation study

The above comparison illustrates the effectiveness of PHIAF, and the success of PHIAF is a result of its design: a GAN-based data augmentation module that generates high-quality pseudo samples and a PHI prediction module that fuses DNA and protein sequence-derived features with an attention mechanism. Here, we conduct an ablation study to elaborate the contribution of these components. We consider the following variants of PHIAF:

- PHIAF-D is a variant that does not use DNA-level features.
- PHIAF-P is a variant that does not use protein-level features.
- PHIAF-A is a variant that does not use the attention layer.
- PHIAF-G is a variant that does not use pseudo samples.

Table 2 shows the results of PHIAF and its four variants under 5-CV. The performance of PHIAF decreases when any component is removed, which means that all the components are critical

Table 2. The performance of PHIAF and different variants using 5-fold cross-validation

	AUPR	AUC	F1	Acc	Sen	Spe
PHIAF-D	0.80	0.85	0.77	0.77	0.79	0.75
PHIAF-P	0.84	0.86	0.79	0.79	0.80	0.79
PHIAF-A	0.82	0.86	0.64	0.73	0.60	**0.88**
PHIAF-G	0.79	0.82	0.73	0.73	0.75	0.71
PHIAF	**0.86**	**0.88**	**0.81**	**0.81**	**0.83**	0.78

Note: The highest value in each column is bold.

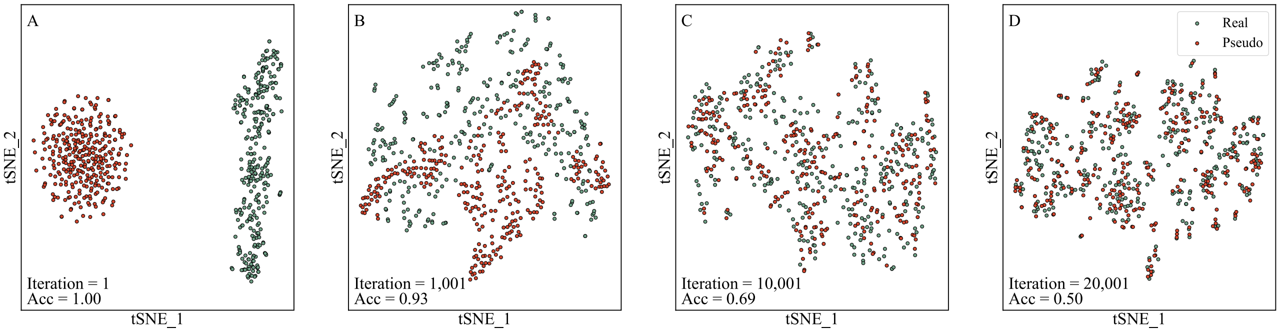

Figure 4. The t-SNE-transformed two-dimensional visualization of real and pseudo samples during different training iterations of GAN (A, 1 iteration; B, 1001 iterations; C, 10 001 iterations; D, 20 001 iterations).

for PHIAF. PHIAF-G suffers the greatest performance decrease, with the AUC and AUPR decreased by 6% and 7%, respectively, followed by PHIAF-D (the AUC and AUPR decreased by 3% and 6%, respectively), and PHIAF-A (the AUC and AUPR decreased by 2% and 4%, respectively). Comparison between the results of PHIAF and PHIAF-G shows that the use of pseudo samples effectively enhances PHI prediction. In addition, the comparison between PHIAF-D and PHIAF-P indicates that the DNA sequence-derived features are more effective for PHI prediction than are the features originating from protein sequences. Removing the attention layer also leads to poorer performance, indicating that the differences in features must be taken into account.

Discussion

The ablation study shows that the main components of PHIAF make important contributions to PHI prediction. Further, we analyze PHIAF from three aspects.

To demonstrate that the real and pseudo samples are indistinguishable and that the pseudo samples can be used as training positive samples, we use t-distributed stochastic neighbor embedding (t-SNE) [53] to visualize the 2D distribution of real and pseudo positive samples at different training iterations of GAN (Figure 4). In the 1st iteration (Figure 4A), the real samples are distributed far from the pseudo samples, leading to a LOOCV accuracy of 1.00, which suggests that the generator has not learned the distribution of real samples. As shown in Figure 4B, the generator begins to capture the characteristics of real samples, and the discriminator cannot fully distinguish between real and pseudo samples after 1000 iterations (LOOCV accuracy of 0.93). Then, the distributions continue to converge gradually (Figure 4C, LOOCV accuracy reaching 0.69). After 20 000 iterations, the generator and discriminator reach equilibrium, and the real and pseudo samples follow similar distributions (Figure 4D, LOOCV accuracy of 0.50). Thus, high-quality pseudo positive samples are generated for data augment.

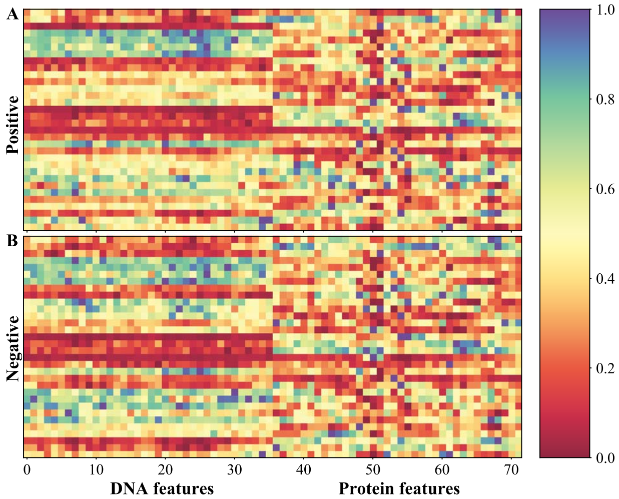

Figure 5. The average weights of positive and negative samples in our dataset in the attention layer (A is positive samples, B is negative samples).

Further, we analyze the weights in the attention layer assigned to different features to investigate the importance of features learned by the attention mechanism. Figure 5A and B shows that the attention weight distributions of positive and negative samples are similar, which indicates that some features play the same important role in positive and negative samples. Additionally, we compare the attention weights in the DNA- and protein-level, where protein-level features are generally assigned lower weights than are DNA-level features. These results confirm that the DNA-level features are more important than protein-level features for PHI prediction and that the attention layer in the CNN not only enhances the prediction performance but also effectively assigns weights based on the importance of different features.

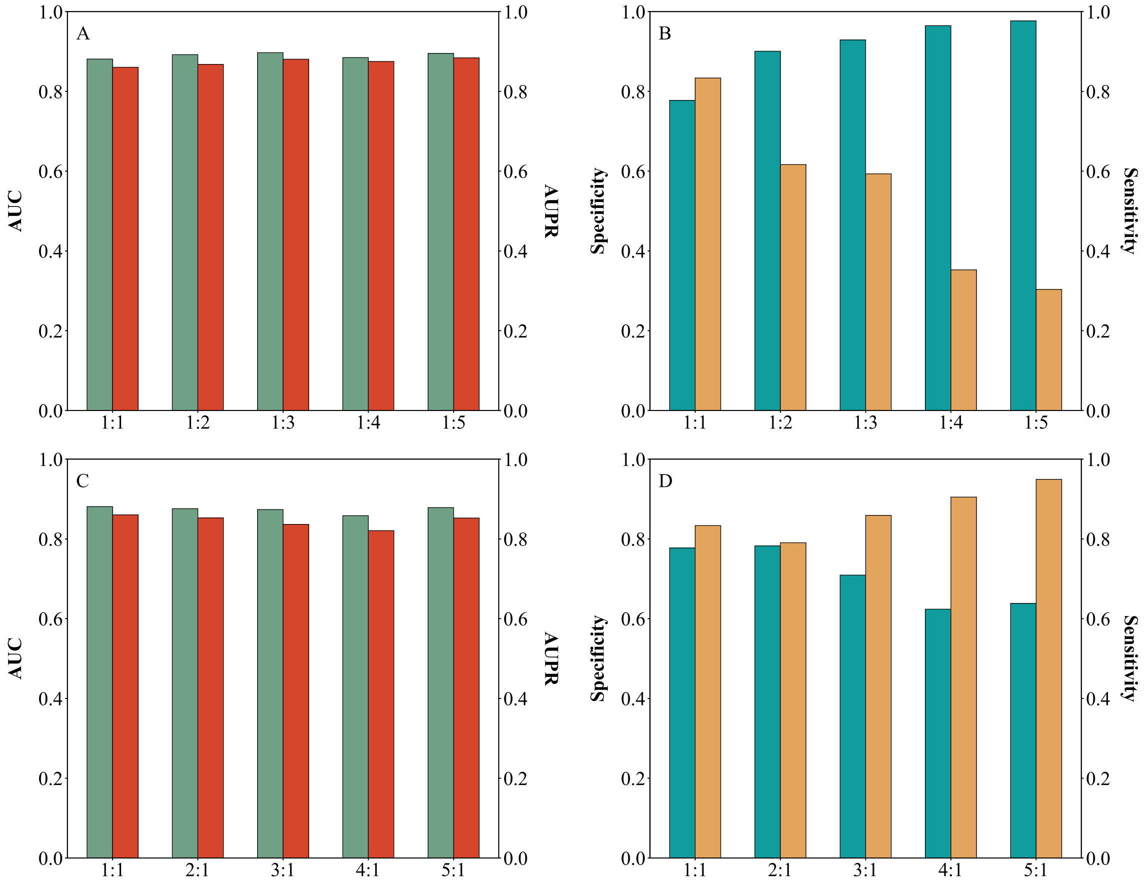

Figure 6. The performance of models based on datasets with different ratios of positive and negative samples under 5-CV (x-axis represents different ratios).

Finally, we build PHIAF models based on datasets with different ratios of positive and negative samples (1:1, 1:2, 1:3, 1:4, 1:5, 2:1, 3:1, 4:1 and 5:1) to analyze how the number of negative or pseudo positive samples influences the performance of PHIAF. As shown in Figures 6A and 6C, PHIAF produces similar AUC and AUPR scores (vary within 2%) on datasets with different data imbalance ratios. As the number of negative samples increases (Figure 6B), PHIAF achieves lower sensitivity (decreased from 0.83 to 0.30) and higher specificity (increased from 0.78 to 0.98). By contrast, PHIAF produces higher sensitivity (increased from 0.79 to 0.95) and lower specificity (decreased from 0.78 to 0.62) when the number of pseudo positive samples increases (Figure 6D). These results suggest that unbalanced datasets failed to significantly enhance the performance of PHI prediction, and samples are likely to be predicted as negatives/positives.

Table 3. The phage-host interactions predicted by PHIAF (these phages and hosts appeared in NCBI after 1 January 2021)

Phages (accession number)	Hosts (accession number)	Evidence
NC_052979	NZ_CP029736	[54]
NC_053009	NZ_CP029736	[54]
NC_052969	NC_003911	[55]
NC_052979	NC_017731	[54]
NC_053009	NC_017731	NA
NC_052979	NC_003911	NA
NC_053009	NC_003911	NA
NC_052969	NZ_CP029736	NA
NC_052969	NC_017731	NA

NA represents this interaction without the evidence reported in literature.

Case study

In this section, we conduct a case study to estimate the ability of PHIAF to predict unknown new PHIs. We first train the PHIAF using the known interactions of all phages and hosts that appeared in NCBI before 1 January 2021, to identify all pairs between the remaining phages and hosts (the remainder represents phages and hosts that appeared in NCBI after 1 January 2021). Then, we rank the PHIs according to the prediction score and search the newly published literature to verify whether the predicted PHI has been confirmed by biological experiments. We list these predicted PHIs in Table 3 (sorted by prediction scores); four of these pairs have been verified by recently published literature. For example, the analysis results in [54] described that Kokobel1 (NCBI accession number: NC_052979) can kill some strains of *Providencia rettgeri* (NCBI accession number: NZ_CP029736) and *Providencia stuartii* (NCBI accession number: NC_017731). Zhan *et al.* [55] reported five bacteriophages infecting *Ruegeria pomeroyi* DSS-3 (NCBI accession number: NC_003911), one of which is vB_RpoS-V16 (NCBI accession number: NC_052969). The results of this case study demonstrate that PHIAF can help to identify novel PHIs and narrow the scope of candidates for further biological experiments.

Conclusion

The overuse of antibiotics has led to several severe challenges in the treatment of bacterial diseases. As one of the most promising alternatives to antibiotics for the treatment of bacterial diseases, phage therapy has received widespread attention. Determining PHIs is extremely important for understanding whether phages can be used to treat bacterial diseases. In the present study, we propose a PHIAF method for PHI prediction that utilizes a GAN to generate high-quality pseudo samples, fuses the features derived from DNA and protein sequences for better performance and uses an attention mechanism to provide interpretability of the prediction model. A comparison with state-of-the-art methods via 5-CV demonstrates that PHIAF achieves the best PHI prediction (performance approximately 13.64% and 14.75% improvement, on average, in terms of AUC and AUPR). Moreover, an ablation study illustrates the contributions of each component of PHIAF, the data augmentation module makes the greatest contribution to the prediction model. Further, a case study is performed to prove the practicable capability of our method. The experimental results indicate that the PHIAF is a promising tool for identifying PHIs.

Despite the good prediction performance of our model, some limitations remain to be addressed. For example, the initial values and ranges of multiple hyper-parameters we set are originated from previous studies and only roughly determined in limited experiments. In the future, we are able to obtain optimal hyper-parameters by training more prediction models. In addition, as a network architecture designed to predict PHIs, PHIAF can be applied to handle other sequence-based classification tasks in bioinformatics.

Key Points

- A generative adversarial network-based data augmentation module is developed to generate high-quality pseudo samples to alleviate the data scarcity problem of phage-host interactions.
- DNA and protein sequence-derived features are combined to effectively improve the phage-host interaction prediction performance.
- The different contributions of DNA and protein features are taken into account through an attention layer, and an attention mechanism provides interpretability of the prediction model.

Supplementary Data

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

Data Availability

The data set and source code can be freely downloaded from https://github.com/mengluli-web/PHIAF or https://github.com/BioMedicalBigDataMiningLab/PHIAF.

Funding

National Natural Science Foundation of China (62072206, 61772381); Huazhong Agricultural University Scientific & Technological Self-innovation Foundation; the Fundamental Research Funds for the Central Universities (2662021JC008).

Acknowledgments

We thank anonymous reviewers for their valuable suggestions.

References

1. Chin CS, Sorenson J, Harris JB, *et al*. The origin of the Haitian cholera outbreak strain. *N Engl J Med* 2010; **364**:33–42.
2. Khan S, Imran A, Malik A, *et al*. Bacterial imbalance and gut pathologies: association and contribution of E. coli in inflammatory bowel disease. *Crit Rev Clin Lab Sci* 2019; **56**:1–17.
3. Khan S. Potential role of Escherichia coli DNA mismatch repair proteins in colon cancer. *Crit Rev Oncol Hematol* 2015; **96**:475–82.
4. Khan S, Zaidi S, Alouffi AS, *et al*. Computational proteome-wide study for the prediction of Escherichia coli protein targeting in host cell organelles and their implication in development of colon cancer. *ACS Omega* 2020; **5**(13): 7254–61.
5. Li J, Zakariah M, Malik A, *et al*. Analysis of Salmonella typhimurium protein-targeting in the nucleus of host cells and the implications in colon cancer: an in-silico approach. *Infect Drug Resist* 2020; **13**:2433–42.
6. Hassel B. Tetanus: pathophysiology, treatment, and the possibility of using botulinum toxin against tetanus-induced rigidity and spasms. *Toxins (Basel)* 2013; **5**:73–83.
7. 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**:30830–43.
8. Davies J, Davies D. Origins and evolution of antibiotic resistance. *Microbiol Mol Biol Rev* 2010; **74**:417–33.
9. Gorski A, Miedzybrodzki R, Wegrzyn G, *et al*. Phage therapy: current status and perspectives. *Med Res Rev* 2020; **40**: 459–63.
10. Kadri SS. Key takeaways from the U.S. CDC's 2019 antibiotic resistance threats report for frontline providers. *Crit Care Med* 2020; **48**:939–45.
11. Cassini A, Högberg LD, Plachouras D, *et al*. Attributable deaths and disability-adjusted life-years caused by infections with antibiotic-resistant bacteria in the EU and the European Economic Area in 2015: a population-level modelling analysis. *Lancet Infect Dis* 2019; **19**:56–66.
12. Towse A, Hoyle CK, Goodall J, *et al*. Time for a change in how new antibiotics are reimbursed: development of an insurance framework for funding new antibiotics based on a policy of risk mitigation. *Health Policy* 2017; **121**: 1025–30.
13. Stokes JM, Yang K, Swanson K, *et al*. A deep learning approach to antibiotic discovery. *Cell* 2020; **180**:688–702.e13.
14. Pires DP, Costa AR, Pinto G, *et al*. Current challenges and future opportunities of phage therapy. *FEMS Microbiol Rev* 2020; **44**:684–700.
15. Edwards RA, McNair K, Faust K, *et al*. Computational approaches to predict bacteriophage-host relationships. *FEMS Microbiol Rev* 2016; **40**:258–72.
16. Villarroel J, Kleinheinz KA, Jurtz VI, *et al*. HostPhinder: a phage host prediction tool. *Viruses* 2016; **8**:116.

17. Liu D, Ma Y, Jiang X, *et al.* Predicting virus-host association by kernelized logistic matrix factorization and similarity network fusion. *BMC Bioinformatics* 2019; **20**:594.
18. Wang W, Ren J, Tang K, *et al.* A network-based integrated framework for predicting virus-prokaryote interactions. *NAR Genom Bioinform* 2020; **2**: lqaa044.
19. Ahlgren NA, Ren J, Lu YY, *et al.* Alignment-free d_2^* oligonucleotide frequency dissimilarity measure improves prediction of hosts from metagenomically-derived viral sequences. *Nucleic Acids Res* 2017; **45**:39–53.
20. 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.
21. Zhang M, Yang L, Ren J, *et al.* Prediction of virus-host infectious association by supervised learning methods. *BMC Bioinformatics* 2017; **18**:60.
22. 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**:5.
23. Mock F, Viehweger A, Barth E, *et al.* VIDHOP, viral host prediction with deep learning. *Bioinformatics* 2021; **37**:318–25.
24. Hauser R, Blasche S, Dokland T, *et al.* Bacteriophage protein-protein interactions. *Adv Virus Res* 2012; **83**:219–98.
25. Alguwaizani S, Park B, Zhou X, *et al.* Predicting interactions between virus and host proteins using repeat patterns and composition of amino acids. *J Healthc Eng* 2018; **2018**:1391265.
26. Young F, Rogers S, Robertson DL. Predicting host taxonomic information from viral genomes: a comparison of feature representations. *PLoS Comput Biol* 2020; **16**:e1007894.
27. Boeckaerts D, Stock M, Criel B, *et al.* Predicting bacteriophage hosts based on sequences of annotated receptor-binding proteins. *Sci Rep* 2021; **11**:1467.
28. Leite DMC, Brochet X, Resch G, *et al.* Computational prediction of inter-species relationships through omics data analysis and machine learning. *BMC Bioinformatics* 2018; **19**: 420.
29. Leite DMC, Lopez JF, Brochet X, *et al.* Exploration of multiclass and one-class learning methods for prediction of phage-bacteria interaction at strain level. In: *International Conference on Bioinformatics and Biomedicine (BIBM)*. Madrid, Spain: IEEE, 2018, 1818–25.
30. 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*. 10.1109/TCBB.2020.3017386.
31. Gao NL, Zhang C, Zhang Z, *et al.* MVP: a microbe-phage interaction database. *Nucleic Acids Res* 2018; **46**:D700–7.
32. Russell DA, Hatfull GF. PhagesDB: the actinobacteriophage database. *Bioinformatics* 2017; **33**:784–6.
33. Mihara T, Nishimura Y, Shimizu Y, *et al.* Linking virus genomes with host taxonomy. *Viruses* 2016; **8**:66.
34. Pruitt KD, Tatusova T, Brown GR, *et al.* NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. *Nucleic Acids Res* 2012; **40**: D130–5.
35. Deng Y, Xu X, Qiu Y, *et al.* A multimodal deep learning framework for predicting drug-drug interaction events. *Bioinformatics* 2020; **36**:4316–22.
36. Li J, Pu Y, Tang J, *et al.* DeepATT: a hybrid category attention neural network for identifying functional effects of DNA sequences. *Brief Bioinform* 2021; **22**: bbaa159.
37. Goodfellow IJ, Pouget-Abadie J, Mirza M, *et al.* Generative Adversarial Nets. In: *Conference on Neural Information Processing Systems (NeurIPS)*. Montreal, Quebec, Canada: Curran Associates, Inc., 2014, 2672–2680.
38. Zhu JY, Park T, Isola P, *et al.* Unpaired image-to-image translation using cycle-consistent adversarial networks. In: *IEEE International Conference on Computer Vision (ICCV)*. Venice, Italy: IEEE, 2017, 2223–32.
39. Wan C, Jones DT. Protein function prediction is improved by creating synthetic feature samples with generative adversarial networks. *Nat Mach Intell* 2020; **2**:540–50.
40. Vaswani A, Shazeer N, Parmar N, *et al.* Attention is all you need. In: *Conference on Neural Information Processing Systems (NeurIPS)*. Long Beach, CA, USA: Curran Associates, Inc., 2017, 5998–6008.
41. Huang Y, Niu B, Gao Y, *et al.* CD-HIT Suite: a web server for clustering and comparing biological sequences. *Bioinformatics* 2010; **26**:680–2.
42. Chen Z, Zhao P, Li F, *et al.* iLearn: an integrated platform and meta-learner for feature engineering, machine-learning analysis and modeling of DNA, RNA and protein sequence data. *Brief Bioinform* 2020; **21**:1047–57.
43. Yi X, Walia E, Babyn P. Generative adversarial network in medical imaging: a review. *Med Image Anal* 2019; **58**:101552.
44. Gao X, Deng F, Yue X. Data augmentation in fault diagnosis based on the Wasserstein generative adversarial network with gradient penalty. *Neurocomputing* 2020; **396**:487–94.
45. Lopez-Paz D, Oquab M. Revisiting classifier two-sample tests. In: *The International Conference on Learning Representations (ICLR)*. Toulon, France: OpenReview.net, 2017.
46. Xu Y, Zhang Z, You L, *et al.* scIGANs: single-cell RNA-seq imputation using generative adversarial networks. *Nucleic Acids Res* 2020; **48**:e85.
47. Wei L, Ye X, Xue Y, *et al.* ATSE: a peptide toxicity predictor by exploiting structural and evolutionary information based on graph neural network and attention mechanism. *Brief Bioinform* 2021;bbab041. 10.1093/bib/bbab041.
48. Xu W, Zhu L, Huang DS. DCDE: an efficient deep convolutional divergence encoding method for human promoter recognition. *IEEE Trans Nanobioscience* 2019; **18**:136–45.
49. Zhang Q, Zhu L, Huang DS. High-order convolutional neural network architecture for predicting DNA-protein binding sites. *IEEE/ACM Trans Comput Biol Bioinform* 2019; **16**:1184–92.
50. Chuai G, Ma H, Yan J, *et al.* DeepCRISPR: optimized CRISPR guide RNA design by deep learning. *Genome Biol* 2018; **19**:80.
51. Zhang S, Zhao L, Zheng CH, *et al.* A feature-based approach to predict hot spots in protein-DNA binding interfaces. *Brief Bioinform* 2020; **21**:1038–46.
52. Tang X, Zhang T, Cheng N, *et al.* usDSM: a novel method for deleterious synonymous mutation prediction using undersampling scheme. *Brief Bioinform* 2021;bbab123. 10.1093/bib/bbab123.
53. LVD M, Geoffrey H. Visualizing data using t-SNE. *J Mach Learn Res* 2008; **9**:2579–605.
54. Rakov C, Ben Porat S, Alkalay-Oren S, *et al.* Targeting biofilm of MDR Providencia stuartii by phages using a catheter model. *Antibiotics* 2021; **10**:375.
55. Zhan Y, Huang S, Chen F. Genome sequences of five bacteriophages infecting the marine Roseobacter bacterium Ruegeria pomeroyi DSS-3. *Microbiol Resour Announc* 2018; **7**:e00959–18.