A Deep Learning-Based Method for Identification of Bacteriophage-Host Interaction

Menglu Li, Yanan Wang, Fuyi Li, Yun Zhao, Mengya Liu, Sijia Zhang, Yannan Bin, A. Ian Smith, Geoffrey I. Webb, Jian Li, Jiangning Song, and Junfeng Xia

Abstract—Multi-drug resistance (MDR) has become one of the greatest threats to human health worldwide, and novel treatment methods of infections caused by MDR bacteria are urgently needed. Phage therapy is a promising alternative to solve this problem, to which the key is correctly matching target pathogenic bacteria with the corresponding therapeutic phage. Deep learning is powerful for mining complex patterns to generate accurate predictions. In this study, we develop PredPHI (Predicting Phage-Host Interactions), a deep learning-based tool capable of predicting the host of phages from sequence data. We collect >3000 phage-host pairs along with their protein sequences from PhagesDB and GenBank databases and extract a set of features. Then we select high-quality negative samples based on the K-Means clustering method and construct a balanced training set. Finally, we employ a deep convolutional neural network to build the predictive model. The results indicate that PredPHI can achieve a predictive performance of 81% in terms of the area under the receiver operating characteristic curve on the test set, and the clustering-based method is significantly more robust than that based on randomly selecting negative samples. These results highlight that PredPHI is a useful and accurate tool for identifying phage-host interactions from sequence data.

Index Terms—Phage-host interaction; bioinformatics; sequence analysis; deep learning; pattern recognition; multi-drug resistance

1 Introduction

BACTERIAL infections have become the greatest challenge in public health care worldwide. In 2014, a World Health Organization report announced that increases in drug resistance were making it more difficult for antibiotics to treat bacterial infections [1]. Furthermore, overconsumption and uncontrolled use of antimicrobials has worsened the situation. Meantime, bacteria are highly adaptable and can rapidly evolve resistance to new antibiotics, thereby significantly reducing drug-related effects [2]. Developing new antibiotics is a time-consuming and cost-ineffective process for pharmaceutical companies; therefore, a practical approach to fighting bacterial infections is urgently needed [3]. Currently, phage therapy is a promising approach that uses viruses to infect and kill bacteria. Upon phages' recognition of specific types of receptors on the bacterial surface, they inject their DNA into the bacteria, resulting in replication, generation of additional phages, and production of an enzyme that dissolved the outer bacterial cell membrane to release the generated phages.

The use of bacteriophages to cure bacterial infections was originally reported in 1919. A French-Canadian microbiologist used them to treat a patient with severe bacillary dysentery [4]. Recent studies confirm the effect of phage therapy as an alternative to antibiotics [5, 6], and many countries have revitalized the interest of phage therapy [7]. An increasing number of studies focus on the mechanism by which phages attack bacteria and the process by which bacteria evolve to defend against phage invasion [8]. The success of phage therapy relies on correctly matching a bacterium to a therapeutic phage. Nevertheless, phage infectivity can vary depending on the host bacterial species and potentially differ in the same strain [9], because both bacteria and bacteriophages regularly adjust their defense and attack mechanisms, respectively [10]. Phage therapy is often conducted using phage cocktails comprising of multiple phage types, including lytic phages to rupture bacterial cells and lysogenic phages to prevent drug resistance, to enable the use of standard antibiotics to treat patients.

Currently, many experimental methods, such as spot

- *M. Li, Y. Bin and J. Xia are with the Key Laboratory of Intelligent Computing and Signal Processing of Ministry of Education, Institutes of Physical Science and Information Technology, and the School of Computer Science and Technology, Anhui University, Hefei, Anhui 230601, China. Email: jfxia@ahu.edu.cn (Corresponding author).*
- *Y. Wang and F. Li are with the Biomedicine Discovery Institute and Department of Biochemistry & Molecular Biology, and Monash Centre for Data Science, Faculty of Information Technology, Monash University, Melbourne, VIC 3800, Australia.*
- *Y. Zhao and A. Smith are with the Biomedicine Discovery Institute and Department of Biochemistry & Molecular Biology, Monash University, Melbourne, VIC 3800, Australia.*
- *M. Liu and S. Zhang are with the Key Laboratory of Intelligent Computing and Signal Processing of Ministry of Education, Institutes of Physical Science and Information Technology, Anhui University, Hefei, Anhui 230601, China.*
- *G. Webb is with the Monash Centre for Data Science, Faculty of Information Technology, Monash University, Melbourne, VIC 3800, Australia.*
- *J. Li is with the Biomedicine Discovery Institute and Department of Microbiology, the Centre to Impact AMR, Monash University, Melbourne, VIC 3800, Australia.*
- *J. Song is with the Biomedicine Discovery Institute and Department of Biochemistry & Molecular Biology, Monash Centre for Data Science, Faculty of Information Technology, and the Centre to Impact AMR, Monash University, Melbourne, VIC 3800, Australia. Email: Jiangning.Song@monash.edu (Corresponding author).*

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2020.3017386, IEEE/ACM Transactions on Computational Biology and Bioinformatics

assays and microfluidic polymerase chain reaction [11], have been used to identify phage-specific hosts. However, these methods require time-consuming experiments to verify whether there is an interaction between the phage and the host, and currently, it is difficult to verify the increasing data. Recent studies report the use of computational methods to predict phage-host pairs, including HostPhinder [12], VirHostMatcher [13], WIsH [14], LMFH-VH [15], ILMF-VH [16] and Leite's method [17, 18]. Specifically, the HostPhinder tool [12] predicts the host corresponding to phage by searching a known reference phage database for the most genetically similar phage. VirHostMatcher [13] predicts the host of a given virus by finding the host with the largest similarity in oligonucleotide frequency (ONF). WIsH [14] is based on the homogeneous Markov model prediction, and this method runs faster than the VirHostMatcher method. LMFH-VH [15] and ILMF-VH [16] integrate information from three networks (phage-phage network, host-host network, and phage-host interactions), and use neighborhood regularization logistic matrix factorization to predict phage-host interactions, where ILMF-VH adds similar network fusion (SNF) to integrate multiple host information compared to LMFH-VH. In these methods, the accurate determination of genome-sequence similarity is critical. Additionally, Leite et al. [17, 18] compared the performance of traditional machine-learning and ensemble-learning methods for predicting phage-host interactions. They used the protein sequence information of bacteriophage and host to encode. In this method, the negative samples in the training set were randomly selected, which might cause the model results to be unstable.

In this study, we first summarize the data, algorithms, and results of the existing methods. Next, we propose a novel deep-learning method that combines clustering methods to select high-quality negative samples to improve the prediction of phage-host pairs. We collect phage-host pairs along with their respective encoded protein sequences from public databases (i.e. PhagesDB [19] and GenBank [20]). We extract a set of informative features based on the amino acid residue frequency, chemical composition, and molecular weight. Then based on these features, we use the K-Means clustering method to select negative samples from the samples that have not been reported to be positives and build a balanced training set [21, 22]. Finally, we construct a predictor of phage-host interactions based on the deep learning model [23], termed PredPHI (Predicting Phage-Host Interactions). Benchmarking tests show that PredPHI is superior to the state-of-the-art methods, and provides a more stable predictive performance than that of the model trained using randomly selecting negative samples.

2 Materials and Methods

2.1 Benchmark Dataset

To construct the training and test sets, we extracted 3,503 phage-host interactions data with the literature support from two major public databases PhagesDB [19] and GenBank [20] in March 2019. We removed phage and bacteria that lack complete genomes before the conversion of all hosts to the species level. At the same time, we deleted those cases where the protein encoded in the phage and host was a hypothetical protein or the protein was named as "n/a". We used the remaining phage and host to form the interaction pair, where the experimentally verified interaction was used as a positive sample. We finally collected a total of 3,469 interactions. Then we split the entire dataset into a training dataset and an independent dataset according to their curation time. More specifically, the data curated earlier than 2016 was used as the independent test set, while the data collected after 2016 was used for model training.

In terms of the negative samples, an ideal situation would be that the negative dataset contained experimentally confirmed non-interacting phage-host pairs. However, to the best of our knowledge, this information is not provided by any currently available data source. Therefore, we used the K-Means clustering method to select highly reliable negative samples from all negative samples, and the number of selected negatives was consistent with that of the positive samples. Then we merged the positive and negative samples to form the training dataset (Training-KM). In addition, a reference dataset was also curated by randomly selecting the negative samples (sampling without the replacement strategy, Training-RS) (Table 1). In order to evaluate the model performance in the real-world, we randomly constructed 10 test sets. Each time, we selected negative samples from the all negative sample set consistent with the number of positive samples in the test set. A statistical summary of each dataset is shown in Table 1. In the K-Means clustering method, we selected the number of initial cluster center points based on the sum of the squared errors (SSE) [24]. We chose the data points that could minimize the SSE as the number of cluster center points (the elbow method). The relationship between SSE and the number of cluster center points is given in Appendix Fig. S1. We finally selected 23 clustering center points and 123 pairs of interactions from each cluster to form a reliable negative sample set.

TABLE 1
Datasets Used in This Study

Dataset	Number of positive samples	Number of negative samples
Training-KM	2,851	2,829
Training-RS	2,851	2,851
Test	618	618

KM represents selected negative samples based on the K-Means clustering method in the data set, and RS represents selected negative samples based on the random selection method.

2.2 Feature Encoding

The interaction between phage and host depends on the ligands and receptors on their surfaces [25]. Therefore, the information that is contained within phage- and host-encoded protein sequences might help predict phage-host interaction [26-29]. In this study, we used phage and host protein sequences to extract informative features.

Using the primary sequence of each protein, we extracted the following 27 features [30-32], which included

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2020.3017386, IEEE/ACM Transactions on Computational Biology and Bioinformatics

the percentage of each amino acid (21 features; 20 amino acids plus one to indicate unknown amino acids, AAC); the abundance of each chemical component in the sequences (5 features; carbon, hydrogen, oxygen, nitrogen, and sulfur, AC); and the molecular weight of the protein (1 feature, MW). The formula is as follows:

$$AAC = N_t/N\,, t \in \{A, C, D, \dots, *\} \quad (1)$$

where N_t represents the number of occurrences of the amino acid t in the protein sequence, while N represents the total number of amino acids in the sequence. AC can be formulated as follows:

$$AC = N_c\,,\ \ c \in \{CHONS\} \quad (2)$$

where N_c represents the number of occurrences of carbon (C), hydrogen (H), oxygen (O), nitrogen (N), and sulfur (S) in the sequence. MW can be calculated as follows:

$$MW = sum(w_t) - (m-1) * 18.01 \quad (3)$$

where w_t represents the molecular weight of the amino acid t, $sum(\cdot)$ is the sum of the molecular weights of all amino acids in the sequence, and m is the length of the sequence.

Therefore, based on the features of the protein sequence, we extracted the representative features of the phage and the host. Each phage (or host) contained multiple proteins. We combined these protein features as the features of the phage (host). The combined forms included: mean, standard deviation (std), maximum (max), minimum (min), median, and variance (var). Then we integrated the features of the phage and host as the final features of the phage-host interaction in the form of 6 × 27 × 2, where "6" denotes six combinations, "27" represents protein sequence features, and "2" represents the bacteriophage and host, respectively. Finally, we reshaped the feature matrix to a one-dimensional vector as the input to the K-Means clustering selection method.

2.3 Deep Learning Model

As a powerful machine learning technique, deep learning has achieved competitive performance on a variety of problems, including natural language processing [33], speech recognition [34], drug-drug interaction [35, 36], protein-DNA binding modeling [37-40], guide RNA designing [41], and human promoter feature extraction [42]. For the deep learning model in this study, as our input feature dimension was 6×27×2, it is impossible to use a convolution kernel with too high dimensionality. The convolution layer of our model used a 3×3 convolution kernel. The pooling layer is often connected after the convolution layer; we selected the max-pooling layer to perform feature extraction after the convolution layer. To reduce the risk of overfitting, we added a dropout layer and set the loss rate to 0.5 [43]. Then we used a fully connected layer (1,024 linear units) to stretch our features. We further connected a dropout layer (with the loss rate of 0.5) after the fully connected layer to reduce the possibility of overfitting in our model. Finally, we used the fully connected layer to generate the output of the predicted probability value. We used a grid search to adjust the learning rate (the range was [0.01, 0.001, 0.0001]) and batch size (the range was [32, 64, 128]) parameters of the convolutional neural network. The final optimal learning rate was 0.0001 and the batch size was 32. The deep learning model was implemented using the Keras deep learning library [44] (Fig. 1).

2.4 Performance Metrics and Cross-validation

To assess the performance of the developed method for predicting phage-host interactions, we employed some commonly used statistical measures [45-49], including Specificity (Spe), Sensitivity (Sen), Recall, Precision (Pre), F1-score (F1), Accuracy (Acc), and Matthews' correlation coefficient (MCC). These performance measures can be calculated as follows:

$$Spe = TN/(TN + FP) \quad (4)$$

$$Sen = Recall = TP/(TP + FN) \quad (5)$$

$$Pre = TP/(TP + FP) \quad (6)$$

$$F1 = 2 * Recall * Pre/(Recall + Pre) \quad (7)$$

$$Acc = TP + TN/(TP + TN + FP + FN) \quad (8)$$

$$MCC = (TP * TN - FP * FN) \Big/ \sqrt{(TP + FN)(TP + FP)(TP + FP)(TP + FP)} \quad (9)$$

where *TP* (true positive) represents the number of positive samples correctly classified in the prediction, *FP* and *FN* (false positive and negative, respectively) denote the numbers of positive and negative samples incorrectly classified, respectively, and *TN* (true negative) represents the number of negative samples correctly classified by the predictor. Furthermore, we also used the area under the receiver-operating characteristic curve (AUC) and the area under the precision-recall curve (AUPR) to assess the overall performance of the model.

3 Results and Discussion

3.1 Overview of Computational Approaches for Predicting Phage-Host Interaction Pairs

Edwards et al. [50] summarized and assessed the experimental and statistical approaches used for identifying phage-host relationships, including occurrence profiles, genetic homology, analysis of CRISPR spacers, exact matches, and similarities in oligonucleotide profiles. Evaluation of sequence homology tends to be the most effective method for identifying known phage-host pairs. Additionally, compositional and abundance-based methods contain significant signals for phage-host classification and provide opportunities to analyze unknowns in viral metagenomes. However, the main limitation of these approaches is the requirement for a sufficiently large number of relatively homogeneous samples. Therefore, to reduce the need for large amounts of data and increase predictive accuracy, other researchers developed similarity network and machine-learning methods to predict phage-host interactions. Table 2 provides a summary of current methods for phage-host interaction prediction, which includes their major data source, algorithms, and results. These methods can be grouped into two major categories in terms of their methodologies: the first group is similarity network-based methods (HostPhinder [12], WIsH [14], VirHostMatcher [13], LMFH-VH [15] and ILMF-VH [16]) and the second

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2020.3017386, IEEE/ACM Transactions on Computational Biology and Bioinformatics

group is machine learning-based methods [17, 18].

3.1.1 Methods Based on a Similarity Network

The similarity network-based methods mainly rely on the assumption that similar phages usually share similar hosts. These methods firstly obtained the sequence information of the phage and host, then calculate the similarity of the phage (host), and finally used different methods to predict the possibility of phage-host interaction. The major difference between different similarity network-based tools is the similarity calculation method and prediction algorithm.

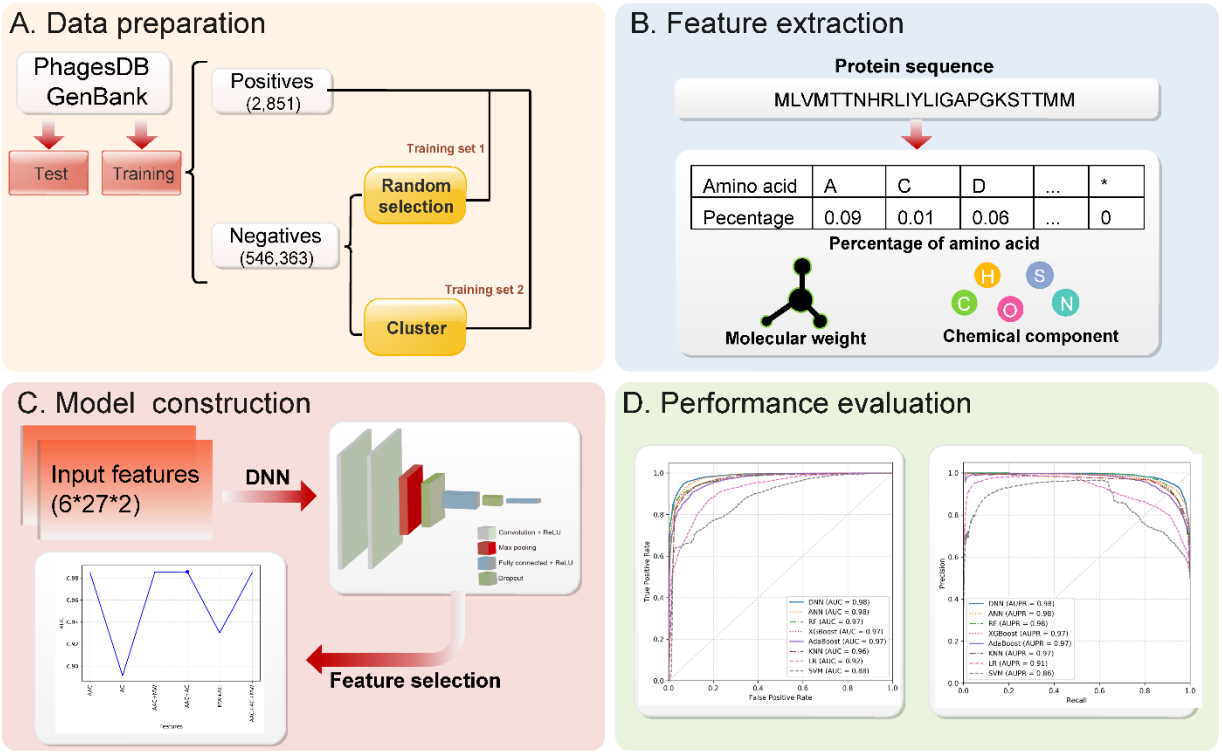


Fig. 1. Flowchart of describing PredPHI prediction of phage-host interactions. A. Download data from PhagesDB and GenBank, divide it into a training set and independent test set, use two methods to select negative samples (the same number of positive and negative samples) to construct training set separately. B. Exact sequence features of the proteins encoded by the phage and host (the features include AAC, AC, and MW). C. Select the optimal classifier (DNN) and the optimal feature subset for the training set. D. Compare the performance of different classifiers.

HostPhinder [12] collected phage and host sequences from Phages.ids-VBI mirrors page (http://mirrors.vbi.vt.edu/mirrors/ftp.ncbi.nih.gov/genomes/IDS/Phages.ids), NCBI viral Genome Resource (https://www.ncbi.nlm.nih.gov/genome/viruses/), EMBL EBI (https://www.ebi.ac.uk/genomes/phage.html), PhagesDB, and GenBank databases. Then based on the sequence information, the author uses the number of co-occurring k-mers (DNA sequences of length k) to compute similarity and inferring the host with the highest similarity measure from the host of the reference phage.

VirHostMatcher [13] downloaded phage and host data from NCBI and used a variety of methods to calculate the ONF similarity, such as Euclidean distance, Manhattan distance, and Chebyshev distance, then predicted the host of a given bacteriophage-based on the host with the greatest ONF similarity. The final experimental results show that the d_2^* is optimal, and the comparison results show that the VirHostMatcher is superior to virus-host abundance covariation, sequence homology to host genomes, and analysis of CRISPR sequences [50].

WIsH [14] improved the accuracy and computational speed based on VirHostMatcher. This method downloaded phage and host data from KEGG and RefSeq, then trained eight potential homogeneous Markov models of each potential host genome by the maximum likelihood, and determined the most likely host. Compared with VirHostMatcher, this method could be run hundreds of times faster.

LMFH-VH [15] obtained the data from VirHostMatcher and the Edwards paper. The authors integrated three network information into a bipartite graph (phage-phage, host-host, and phage-host), where the similarity between the phages was calculated by k-mers. The host similarity was based on the known phage-host interactions and calculated using the Gaussian interaction profile kernel, then predicted unknown interactions by domain regular logic matrix factorization. Finally, the performance of this method was compared with a set of network models for predicting drug-targets on different data sets and was shown to be superior to these methods.

The optimized version based on LMFH-VH is ILMF-VH [16]. ILMF-VH combined three network information mentioned above to form a phage-host heterogeneous network

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2020.3017386, IEEE/ACM Transactions on Computational Biology and Bioinformatics

and used SNF to integrate multiple host information. Compared with the LMFH-VH method, the prediction performance achieved by ILMF-VH was not significant (the host prediction accuracy only increased by 0.49%).

3.1.2 Machine-learning Methods

Similarity network-based methods were mainly depending on known phage-host interactions, which creates a bias toward predicting already known interactions. To eliminate this bias, Leite et al. [17, 18] developed machine-learning methods to predict phage-host interactions. This method initially collected phage- and host-genome sequences from PhagesDB and GenBank, followed by the extraction of various features, including amino acid frequencies, chemical compositions, and molecular weights of each phage- or host-encoded protein. And the author calculated the mean and std of the features across all proteins. Since there are fewer known interactions (positive samples), all those without experimental verification are considered as unknown interactions (negative samples). This method randomly selected negative samples subset from all negative samples, then combined them with positive samples to construct a balanced data set. The model built by traditional machine-learning methods, and a grid search is used to select the optimal parameters. Leite et al. compared the performance of different classifiers, including Random Forest (RF), Support Vector Machines (SVM), Artificial Neural Networks (ANN), K-Nearest Neighbors (KNN), Naïve Bayes (NB), Logistic Regression (LR), and the ensemble-learning models based on these classifiers (Bagging (RF, ANN, KNN), Bagging (RF, SVM, KNN)). The final ensemble-learning model performances were better than a single classifier, and the ensemble-learning model mainly depends on RF to improve performance. The optimal model accuracy was 95.70%.

This method mainly used F1 to evaluate the performance of the model, but F1 is greatly affected by the threshold. Leite et al. do not test the performance of the model on an independent test set, and the method of balancing the data set is to copy a part of the original sample and randomly select a subset of the samples of most categories. This approach leads to overfitting of the model and unstable results. Meanwhile, the prediction performance of the model needs to be improved.

TABLE 2
A Comprehensive Summary of the Reviewed Predictors for Phage-Host Interactions in This Study

Method category	Tool	Year	Major data source	Algorithm	Host level	Accuracy	Code	Code language
Similarity network	Host-Phinder [12]	2016	NCBI, EMBL EBI, PhagesDB, PhAnToMe, GenBank	Genetic similarity	Species, Genus	81%	http://cge.cbs.dtu.dk/services/Host-Phinder	Python and Bash
	VirHost-Matcher [13]	2017	NCBI	Genomic oligonucleotide frequencies	Genus, Family, Order, Class, Phylum, Domain	64%	https://github.com/jessieren/Vir-Host-Matcher	Python and C++
	WIsH [14]	2017	KEGG, RefSeq Virus	Homogeneous markov model	Genus, Family, Order, Class, Phylum	63%	https://github.com/soedinglab/wish	C++
	LMFH-VH [15]	2018	VirHost-Matcher	Logistic matrix factorization	Species	63.17%	https://github.com/liu-dan111/LMFH-VH.git	Python
	ILMF-VH [16]	2019	VirHost-Matcher	Logistic matrix factorization	Species	63.66%	-	-
Machine learning	Leite et al. [17, 18]	2018	Genbank, PhagesDB	Traditional machine learning, ensemble learning	Species	95.70%	https://drive.switch.ch/index.php/s/uo-BpjvY6dnxzrAf	Python

3.2 Model Selection

The current methods are low accuracy, and the model is unstable, we developed our approach to improving the prediction accuracy of the model and the robustness of the model. Firstly, we selected negative samples based on the K-Means clustering method (ensure that the ratio of positive and negative samples is 1) and built a balanced training set (Training-KM). Then based on the Training-KM dataset, we used different classifiers for training and choose an optimal model. The results of the ten-fold cross-validation are shown in Table 3 and Fig. 1D. It could be seen that the DNN (Deep Convolutional Neural Network, the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2020.3017386, IEEE/ACM Transactions on Computational Biology and Bioinformatics

framework described in section 2.3) is optimal on multiple performance metrics.

As can be observed from Fig. 1D and Table 3, the AUC, AUPR, Spe, Pre, and MCC of DNN were slightly higher than the ANN. At the same time, the Sen of DNN was 3% higher than XGBoost (eXtreme Gradient Boosting tree). Compared with other traditional machine-learning methods, the AUC of DNN was 1% higher than RF and AdaBoost, 2% higher than KNN, 6% higher than LR, and 10% higher than the SVM. We finally chose the classifier with the highest AUC (DNN) as the final model.

TABLE 3
Comparative Performance of Different Classifiers in Training-KM Dataset (Ten-Fold Cross-Validation)

	Sen	Spe	Pre	F1	MCC	Acc	AUC
DNN	0.93	**0.94**	**0.94**	**0.93**	**0.86**	**0.93**	**0.98**
KNN	**0.94**	0.88	0.88	0.91	0.81	0.91	0.96
SVM	0.76	0.84	0.83	0.79	0.60	0.80	0.88
RF	0.92	0.92	0.92	0.92	0.84	0.92	0.97
ANN	**0.94**	0.92	0.92	**0.93**	0.85	**0.93**	**0.98**
LR	0.83	0.85	0.84	0.84	0.68	0.84	0.92
XGBoost	0.90	0.93	0.93	0.92	0.84	0.92	**0.98**
AdaBoost	0.90	0.89	0.89	0.90	0.80	0.90	0.97

The highest value in each column is bold.

3.3 The Effect of Different Features on Predictive Performance

We assessed the performance of six different combinations of three types of features (AAC, AC, AAC+MW, AAC+AC, MW+AC, AAC+AC+MW) using DNN based on ten-fold cross-validation (Fig. 2). The dataset used is Training-KM. The MW feature is not used alone because MW is only one dimension and cannot be input into the classifier for training and test. The AC feature build model result is poor, while the AAC feature's model is better, which shows that the amino acid features generated using protein sequence could improve model performance more than chemical element features in the model building process. At the same time, the different feature combinations models were superior to the models constructed from a single feature. This indicates the necessity and importance of developing effective combinations of feature types, rather than using a single feature type alone.

The performance of the model constructed by the AAC and AC features combination was better than the model built by the AAC+AC+MW features. This shows that adding the MW feature based on AAC+AC is a negative impact on the prediction of the model. It may because the MW feature and the AAC feature are mutually exclusive features, which causes noise in the model construction process and affects the prediction performance of the model. The reason for this speculation is to compare the results of the combination of the AAC feature construction model and AAC+MW, and the results of the AC feature construction model and AC+MW. Only use the AAC feature construct model result is slightly higher than the AAC+MW, but the performance of the model built by AC was significantly worse than the performance of the model built by AC+MW. It suggests that the combination of AC and MW can improve the performance of the model in this experiment, while the combination of AAC and MW reduces the performance of the model. Based on all performance metrics, we finally chose the combination of features of AAC and AC to construct a model.

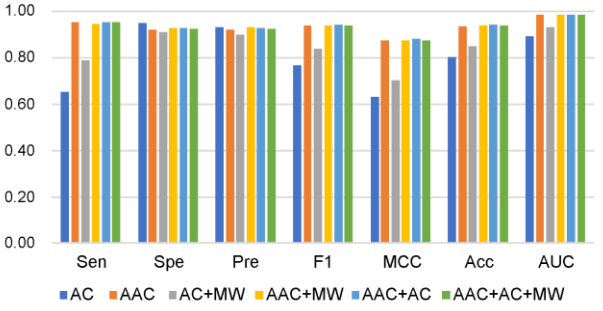
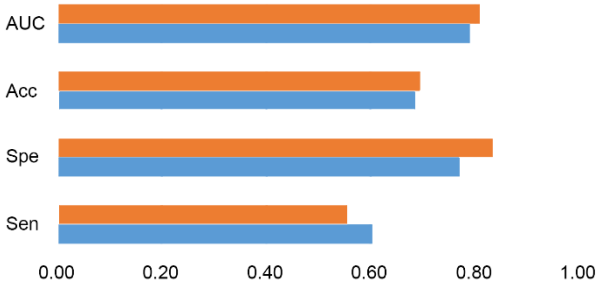


Fig. 2. Comparative importance of different feature combinations.

3.4 The Effect of Different Negative Sample Selection Methods

Based on the previously determined optimal feature combination (AAC+AC), and the optimal classifier (DNN), we constructed our prediction model (PredPHI). To assess the effectiveness of the proposed strategy for selecting negative samples, we experimented on ten independent test sets and averaged the performance result on these ten sets. There was not much difference between the ten test results. The control experiment used a random selection of negative samples. Two training sets are Training-KM and Training-RS datasets. Based on the above two datasets, we used the same features and classifier to train the model, and the final comparison results are shown in Fig. 3. It can be observed from Fig. 3 that the AUC value of the model trained on the Training-KM dataset was 2% higher than that on the Training-RS dataset. It is worth mentioning that we have also constructed 10 models based on the random negative selection method and still got the similar performance (data not shown). Overall, the model based on a selection of negative samples using the clustering method was superior to the random negative selection method in terms of the model robustness and performance.



	Sen	Spe	Acc	AUC
K-Means	0.56	**0.83**	**0.69**	**0.81**
Random	**0.60**	0.77	**0.69**	0.79

Fig. 3. Comparative performance of two different negative sample selection methods in the test set.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2020.3017386, IEEE/ACM Transactions on Computational Biology and Bioinformatics

3.5 Performance Evaluation of Different Machine Learning Methods

By using the Training-KM dataset and the optimal feature set (AAC+AC), we built a model PredPHI based on DNN and compare their performance with different classifiers (ANN, KNN, RF, NB, SVM, and LR) in the test set. Fig. 4 shows the performance comparison of the ROC and PR curves of different machine learning classifiers. From Fig. 4A and B, we can see that the PredPHI was the best performance, followed by the ANN model. The AUC of our model was 1% higher than the ANN model, and 8%, 15%, 18%, 22%, and 25% higher than KNN, NB, RF, SVM, and LR, respectively. the AUPR of our model was 4% higher than the ANN model. The prediction performance ranking is the same as the model prediction in the ten-fold cross-validation experiment discussed above. The PredPHI is the best model, followed by the ANN model, and then other traditional machine-learning methods. It shows that the model constructed by the neural network framework is superior to traditional machine-learning methods. And the method of a convolutional neural network was superior to conventional neural networks (multi-layer perceptron). Overall, the results of Fig. 4 show that the use of a deep convolutional neural network could indeed improve the model performance.

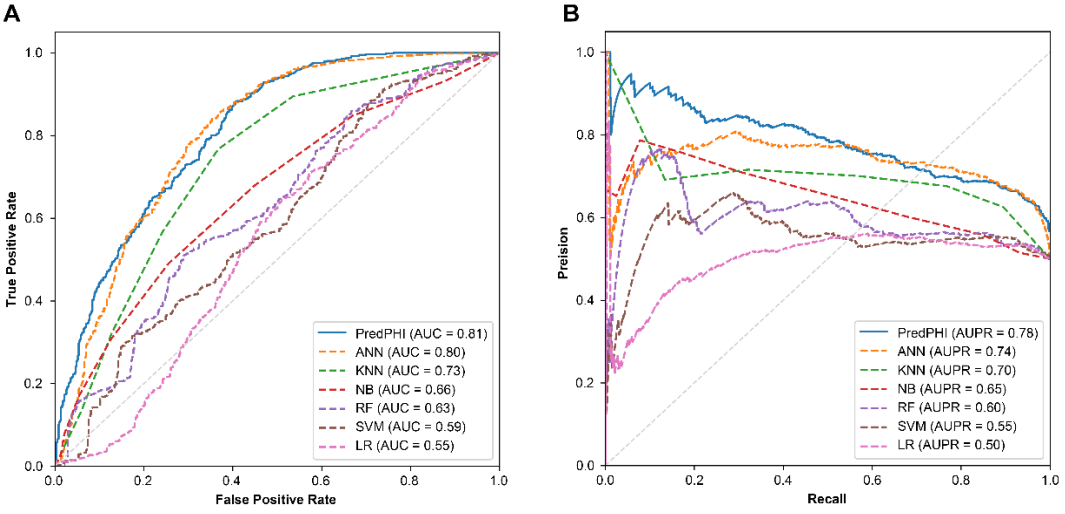
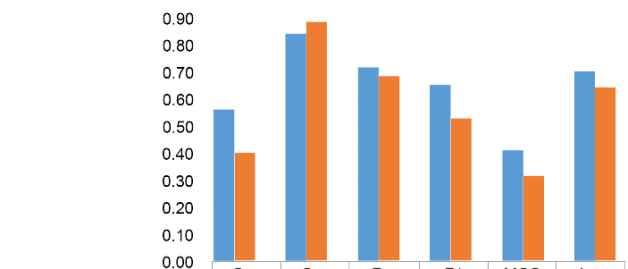


Fig. 4. Comparative performance of PredPHI with different machine-learning methods in the test set.

3.6 Performance Evaluation of the State-of-the-art Methods

The existing methods for predicting phage-host interaction can be divided into two categories: the first category is based on the calculation of the similarity between the phage and host, which relies on the existing phage-host interactions data available in the public databases. It cannot be applied to the independent test set constructed by the phage and host for which neither appeared in the training set. In this section, we compared the second method in the test set. The second category [24,25] randomly selects negative samples, and use features (such as the combination of AAC, AC and MW features) to train the models based on different traditional machine-learning and ensemble-learning methods. The final optimal model is the ensemble-learning method (Bagging (RF, ANN, KNN)). This ensemble-learning method used a hard-voting strategy; as such, the AUC value could not be obtained. Therefore, we compared this method based on other performance metrics. We trained these base classifiers separately on the Training-RS dataset and selected the optimal parameters. The final compare results are shown in Fig. 5.

In Fig. 5, the Sen, Pre, F1, MCC, and Acc of PredPHI was 16%, 3%, 12%, 11%, and 5% higher than the Bagging (RF, ANN, KNN) method, Spe was 5% lower than the Bagging method. It shows that the result of the model was superior to the state-of-the-art methods. The possible reasons are: first, PredPHI uses the K-Means clustering to select representative negative samples, which is beneficial for the model training. Second, the features used were AAC+AC, according to section 3.2. We found that adding MW features would reduce the model performance. Therefore, compared with the features used in the state-of-the-art method (AAC+AC+MW), the AAC+AC features we used could effectively improve our model performance. Finally, the use of a deep convolutional neural network further helped to improve the model performance.



	Sen	Spe	Pre	F1	MCC	Acc
PrePHI	**0.56**	0.83	**0.71**	**0.64**	**0.40**	**0.69**
Bagging (RF, ANN, KNN)	0.40	**0.88**	0.68	0.52	0.31	0.64

Fig. 5. Comparative performance of PredPHI with a previous ensemble learning method in the test set.

4 Conclusion

Effective phage treatment can potentially address the continued emergence of antibiotic resistance. However, the main challenge of effective phage treatment is the rapid and accurate identification of correct phage-target interactions. In this study, we offered a comprehensive survey of existing computational methods employed to predict phage-host interactions and developed a deep-learning technique for this purpose. The benchmark results demonstrate that the use of specific phage- and host-derived features could greatly enhance the predictive power of the model. Moreover, we used the K-Means clustering method to select representative negative samples and keep a consistent number of positive samples to build a balanced training set. The results suggest that this clustering-based method of selecting negative samples led to a more stable performance than the method of randomly selecting negative samples. Besides, the AAC and AC features were found to be particularly useful for predicting phage-host interactions. Taken together, the results indicate that our method achieved superior performance compared with the existing methods in terms of performance metrics on the independent test set, highlighting its usefulness for accurate prediction of phage-host interactions. In view of the increase in the volume of available phage-host data, the application of this method could aid the development of personalized treatment for bacterial infections.

One major drawback of this study is that the deep learning model is not interpretable and difficult for users to understand. Deep learning models are often regarded as 'black boxes'; although the complex architecture of deep learning models improves the predictive performance, this also makes such models difficult to interpret. Due to the balance between accuracy and interpretability, complex models with a better performance may tend to be more difficult to interpret than simple models. In the future, we will attempt to improve the interpretability of the model while maintaining its predictive performance. We will integrate the sequence similarity information of the phage and host to select negative samples and determine higher quality negative samples [51], and examine the possibility of combining the sequence and structural information [52] of the proteins encoded respectively by the phage and the host as the representative features of the phage-host interaction pair and constructing a more robust and accurate model. In addition, the current PredPHI method relies on the phage- and host-encoding protein sequences to extract features, thereby preventing us from predicting phages and hosts that contain non-coding genomic regions. To address this, we will develop the model based on the original genomic sequence data of bacteriophages and hosts in our future work. The source codes of PredPHI are freely accessible at Github (https://github.com/xialab-ahu/PredPHI).

Acknowledgments

This work was financially supported by the National Natural Science Foundation of China (61672037, 21601001, 11835014 and U19A2064), the Anhui Provincial Outstanding Young Talent Support Plan (gxyqZD2017005), the Young Wanjiang Scholar Program of Anhui Province, the Recruitment Program for Leading Talent Team of Anhui Province (2019-16), the China Postdoctoral Science Foundation Grant (2018M630699), the Anhui Provincial Postdoctoral Science Foundation Grant (2017B325), and the Key Project of Anhui Provincial Education Department (KJ2017ZD01). JL and JS's work was supported by grants from the National Health and Medical Research Council of Australia (NHMRC) (1144652 and 1127948), the National Institute of Allergy and Infectious Diseases of the National Institutes of Health (R01 AI111965) and a Major Inter-Disciplinary Research (IDR) Grant Awarded by Monash University.

References

[1] D. I. Andersson, "Persistence of antibiotic resistant bacteria," *Curr Opin Microbiol.,* vol. 6, no. 5, pp. 452-456, 2003.

[2] R. Laxminarayan, A. Duse, C. Wattal, A. K. M. Zaidi, H. F. L. Wertheim, N. Sumpradit, E. Vlieghe, G. L. Hara, I. M. Gould, H. Goossens, C. Greko, A. D. So, M. Bigdeli, G. Tomson, W. Woodhouse, E. Ombaka, A. Q. Peralta, F. N. Qamar, F. Mir, S. Kariuki, Z. A. Bhutta, A. Coates, R. Bergstrom, G. D. Wright, E. D. Brown, and O. Cars, "Antibiotic resistance-the need for global solutions," *Lancet Infect Dis.,* vol. 13, no. 12, pp. 1057-1098, 2013.

[3] G. W. Hanlon, "Bacteriophages: an appraisal of their role in the treatment of bacterial infections," *Int J Antimicrob Agents.,* vol. 30, no. 2, pp. 118-128, 2007.

[4] D. Harper, J. Anderson, and M. Enright, "Phage therapy: delivering on the promise," *Ther Deliv.,* vol. 2, no. 7, pp. 935-947, 2011.

[5] R. Capparelli, M. Parlato, G. Borriello, P. Salvatore, and D. Iannelli, "Experimental phage therapy against Staphylococcus aureus in mice," *Antimicrob Agents Chemother.,* vol. 51, no. 8, pp. 2765-2773, 2007.

[6] A. Wright, C. Hawkins, E. Änggård, and D. Harper, "A controlled clinical trial of a therapeutic bacteriophage preparation in chronic otitis due to antibiotic-resistant pseudomonas aeruginosa; a preliminary report of efficacy," *CClin Otolaryngol.,* vol. 34, no. 4, pp. 349-357, 2009.

[7] S. Reardon, "Phage therapy gets revitalized," *Nature.,* vol. 510, no. 7503, pp. 15, 2014.

[8] K. D. Seed, "Battling phages: how bacteria defend against viral attack," *PLoS Pathog.,* vol. 11, no. 6, pp. e1004847, 2015.

[9] C. O. Flores, J. R. Meyer, S. Valverde, L. Farr, and J. S. Weitz, "Statistical structure of host-phage interactions," *Proc Natl Acad Sci USA.,* vol. 108, no. 28, pp. E288-E297, 2011.

[10] J. E. Samson, A. H. Magadán, M. Sabri, and S. Moineau, "Revenge of the phages: defeating bacterial defences," *Nat Rev Microbiol.,* vol. 11, no. 10, pp. 675, 2013.

[11] A. D. Tadmor, E. A. Ottesen, J. R. Leadbetter, and R. Phillips, "Probing individual environmental bacteria for viruses by using microfluidic digital PCR," *Science,* vol. 333, no. 6038, pp. 58-62, 2011.

[12] J. Villarroel, K. A. Kleinheinz, V. I. Jurtz, H. Zschach, O. Lund, M. Nielsen, and M. V. Larsen, "HostPhinder: a phage host prediction tool," *Viruses,* vol. 8, no. 5, pp. 116, 2016.

[13] N. A. Ahlgren, J. Ren, Y. Y. Lu, J. A. Fuhrman, and F. Sun,

"Alignment-free d_2^* oligonucleotide frequency dissimilarity measure improves prediction of hosts from metagenomically-derived viral sequences," *Nucleic Acids Res.*, vol. 45, no. 1, pp. 39-53, 2017.

[14] C. Galiez, M. Siebert, F. Enault, J. Vincent, J. Söding, and I. Birol, "WIsH: who is the host? predicting prokaryotic hosts from metagenomic phage contigs," *Bioinformatics*, vol. 33, no. 19, pp. 3113-3114, 2017.

[15] D. Liu, X. Hu, T. He, and X. Jiang, "Virus-host association prediction by using kernelized logistic matrix factorization on heterogeneous networks," in IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Madrid, Spain, 2018, pp. 108-113.

[16] D. Liu, Y. Ma, X. Jiang, and T. He, "Predicting virus-host association by kernelized logistic matrix factorization and similarity network fusion," *BMC Bioinform.*, vol. 20, no. 16, pp. 594, 2019.

[17] D. M. C. Leite, X. Brochet, G. Resch, Y.-A. Que, A. Neves, and C. Peña-Reyes, "Computational prediction of inter-species relationships through omics data analysis and machine learning," *BMC Bioinform.*, vol. 19, no. Suppl 14, pp. 420, 2018.

[18] D. M. C. Leite, J. F. Lopez, X. Brochet, M. Barreto-Sanz, Y.-A. Que, G. Resch, and C. Peña-Reyes, "Exploration of multiclass and one-class learning methods for prediction of phage-bacteria interaction at strain level," in IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Madrid, Spain, 2018, pp. 1818-1825.

[19] D. A. Russell, and G. F. Hatfull, "PhagesDB: the actinobacteriophage database," *Bioinformatics*, vol. 33, no. 5, pp. 784-786, 2017.

[20] D. A. Benson, M. Cavanaugh, K. Clark, I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, and E. W. Sayers, "GenBank," *Nucleic Acids Res.*, vol. 45, no. D1, pp. D37-D42, 2017.

[21] Y. Zhao, X. Chen, and J. Yin, "Adaptive boosting-based computational model for predicting potential miRNA-disease associations," *Bioinformatics*, vol. 35, no. 22, pp. 4730-4738, 2019.

[22] B. Wang, L. Wang, C. Zheng, and Y. Xiong, "Imbalance data processing strategy for protein interaction sites prediction," *IEEE/ACM Trans Comput Biol Bioinform*, pp. 1-1, 2019.

[23] S. Hu, P. Chen, P. Gu, and B. Wang, "A deep learning-based chemical system for QSAR prediction," *IEEE J Biomed Health Inform*, pp. 1-1, 2020.

[24] R. C. Hunt, V. L. Simhadri, M. Iandoli, Z. E. Sauna, and C. Kimchi-Sarfaty, "Exposing synonymous mutations," *Trends Genet.*, vol. 30, no. 7, pp. 308-321, 2014.

[25] E. D. Coelho, J. P. Arrais, S. Matos, C. Pereira, N. Rosa, M. J. Correia, M. Barros, and J. L. Oliveira, "Computational prediction of the human-microbial oral interactome," *BMC Systems Biology*, vol. 8, no. 1, pp. 24, 2014.

[26] A. Deng, H. Zhang, W. Wang, J. Zhang, D. Fan, P. Chen, and B. Wang, "Developing computational model to predict protein-protein Interaction sites based on the XGBoost algorithm," *Int. J. Mol. Sci.*, vol. 21, no. 7, pp. 2274, 2020.

[27] J.-F. Xia, X.-M. Zhao, and D.-S. Huang, "Predicting protein-protein interactions from protein sequences using meta predictor," *Amino Acids*, vol. 39, no. 5, pp. 1595-1599, 2010.

[28] Z.-H. You, Y.-K. Lei, J. Gui, D.-S. Huang, and X. Zhou, "Using manifold embedding for assessing and predicting protein interactions from high-throughput experimental data," *Bioinformatics*, vol. 26, no. 21, pp. 2744-2751, 2010.

[29] D.-S. Huang, L. Zhang, K. Han, S. Deng, K. Yang, and H. Zhang, "Prediction of protein-protein interactions based on protein-protein correlation using least squares regression," *Curr Protein Pept Sci.*, vol. 15, no. 6, pp. 553-560, 2014.

[30] Z.-H. You, L. Zhu, C.-H. Zheng, H.-J. Yu, S.-P. Deng, and Z. Ji, "Prediction of protein-protein interactions from amino acid sequences using a novel multi-scale continuous and discontinuous feature set," *BMC Bioinform.*, vol. 15, no. 15, pp. S9, 2014.

[31] Z. Chen, P. Zhao, F. Li, A. Leier, T. T. Marquez-Lago, Y. Wang, G. I. Webb, A. I. Smith, R. J. Daly, K.-C. Chou, and J. Song, "iFeature: a python package and web server for features extraction and selection from protein and peptide sequences," *Bioinformatics*, vol. 34, no. 14, pp. 2499-2502, 2018.

[32] Z. Chen, P. Zhao, F. Li, T. T. Marquez-Lago, A. Leier, J. Revote, Y. Zhu, D. R. Powell, T. Akutsu, G. I. Webb, K.-C. Chou, A. I. Smith, R. J. Daly, J. Li, and J. Song, "iLearn: an integrated platform and meta-learner for feature engineering, machine-learning analysis and modeling of DNA, RNA and protein sequence data," *Brief. Bioinformatics*, pp. bbz041, 2019.

[33] T. Young, D. Hazarika, S. Poria, and E. Cambria, "Recent Trends in Deep Learning Based Natural Language Processing," *IEEE Comput Intell Mag*, vol. 13, no. 3, pp. 55-75, 2018.

[34] D. Amodei, S. Ananthanarayanan, R. Anubhai, J. Bai, E. Battenberg, C. Case, J. Casper, B. Catanzaro, Q. Cheng, and G. Chen, "Deep speech 2: end-to-end speech recognition in English and mandarin," in International Conference on Machine Learning (ICML), New York, USA, 2016, pp. 173-182.

[35] Y. Deng, X. Xu, Y. Qiu, J. Xia, W. Zhang, and S. Liu, "A multimodal deep learning framework for predicting drug-drug interaction events," *Bioinformatics*, pp. btaa501, 2020.

[36] J. Y. Ryu, H. U. Kim, and S. Y. Lee, "Deep learning improves prediction of drug–drug and drug–food interactions," *Proc Natl Acad Sci India Sect B Biol Sci*, vol. 115, no. 18, pp. E4304-E4311, 2018.

[37] Q. Zhang, Z. Shen, and D.-S. Huang, "Modeling in-vivo protein-DNA binding by combining multiple-instance learning with a hybrid deep neural network," *Sci Rep.*, vol. 9, no. 1, 2019.

[38] Q. Zhang, L. Zhu, W. Bao, and D.S.Huang, "Weakly-Supervised Convolutional Neural Network Architecture for Predicting Protein-DNA Binding," *IEEE/ACM Trans Comput Biol Bioinform*, vol. 17, no. 2, pp. 679-689, 2020.

[39] Q. Zhang, L. Zhu, and D.S.Huang, "High-order convolutional neural network architecture for predicting DNA-protein binding sites," *IEEE/ACM Trans Comput Biol Bioinform*, vol. 16, no. 4, pp. 1184-1192, 2019.

[40] B. Alipanahi, A. Delong, M. T. Weirauch, and B. J. Frey, "Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning," *Nat. Biotechnol.*, vol. 33, no. 8, pp. 831-838, 2015.

[41] G. Chuai, H. Ma, J. Yan, M. Chen, N. Hong, D. Xue, C. Zhou, C. Zhu, K. Chen, B. Duan, F. Gu, S. Qu, D. Huang, J. Wei, and Q. Liu, "DeepCRISPR: optimized CRISPR guide RNA design by deep learning," *Genome Biol.*, vol. 19, no. 1, pp. 80, 2018.

[42] W. Xu, L. Zhu, and D.-S. Huang, "DCDE: an efficient deep convolutional divergence encoding method for human promoter recognition," *IEEE Trans Nanobioscience*, vol. 18, no. 2, pp. 136-145,

2019.

[43] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *J Mach Learn Res.*, vol. 15, no. 1, pp. 1929-1958, 2014.

[44] A. Gulli, and S. Pal, *Deep learning with keras*: Packt Publishing Ltd, 2017.

[45] Z. Yue, X. Chu, and J. Xia, "PredCID: prediction of driver frameshift indels in human cancer," *Brief. Bioinformatics*, 2020.

[46] W. Zhang, Z. Li, W. Guo, W. Yang, and F. Huang, "A fast linear neighborhood similarity-based network link inference method to predict microRNA-disease associations," *IEEE/ACM Trans Comput Biol Bioinform*, 2019.

[47] F. Shi, Y. Yao, Y. Bin, C. Zheng, and J. Xia, "Computational identification of deleterious synonymous variants in human genomes using a feature-based approach," *BMC Medical Genom.*, vol. 12, no. 1, pp. 12, 2019.

[48] W. Zhang, K. Jing, F. Huang, Y. Chen, B. Li, J. Li, and J. Gong, "SFLLN: a sparse feature learning ensemble method with linear neighborhood regularization for predicting drug–drug interactions," *Inf. Sci.*, vol. 497, pp. 189-201, 2019.

[49] Y. Gong, Y. Niu, W. Zhang, and X. Li, "A network embedding-based multiple information integration method for the MiRNA-disease association prediction," *BMC Bioinform.*, vol. 20, no. 1, pp. 468, 2019.

[50] R. A. Edwards, K. McNair, K. Faust, J. Raes, and B. E. Dutilh, "Computational approaches to predict bacteriophage-host relationships," *FEMS Microbiol Rev.*, vol. 40, no. 2, pp. 258-272, 2016.

[51] W. M. Chen, S. A. Danziger, J. H. Chiang, and J. D. Aitchison, "PhosphoChain: a novel algorithm to predict kinase and phosphatase networks from high-throughput expression data," *Bioinformatics*, vol. 29, no. 19, pp. 2435-2444, 2013.

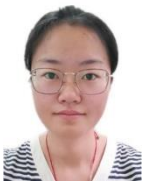
[52] S. Zhang, L. Zhao, C.-H. Zheng, and J. Xia, "A feature-based approach to predict hot spots in protein–DNA binding interfaces," *Briefings in bioinformatics*, 2019.

Menglu Li is currently a master student in the School of Computer Science and Technology and the Institutes of Physical Science and Information Technology, Anhui University. Her research interests are bioinformatics, machine learning, and pattern recognition.

Yanan Wang received his MEng degree from Shanghai Jiao Tong University, China. He is currently a PhD student in the Department of Biochemistry and Molecular Biology and Biomedicine Discovery Institute, Monash University, Australia. His research interests are bioinformatics, computational oncology, machine learning, and pattern recognition.

Fuyi Li is currently a PhD candidate in the Biomedicine Discovery Institute and Department of Biochemistry and Molecular Biology, Monash University, Australia. His research interests are bioinformatics, computational biology, machine learning, and data mining.

Yun Zhao received his Master degree in Information Technology from Monash University, Melbourne, Australia. He is currently a research assistant in the Biomedicine Discovery Institute and the Department of Biochemistry and Molecular Biology, Monash University, Australia. His research interests are bioinformatics, computational oncology, machine learning, and pattern recognition.



Mengya Liu is currently a master student in the Institutes of Physical Science and Information Technology, Anhui University. Her research interests are bioinformatics and machine learning.

Sijia Zhang is currently a master student in the Institutes of Physical Science and Information Technology, Anhui University. Her research interests are bioinformatics and machine learning.

Yannan Bin received the PhD degree in 2013 from Central South University. She is currently a postdoctor in the School of Computer Science and Technology and the Institutes of Physical Science and Information Technology, Anhui University, China. Her research interests are bioinformatics and machine learning.

A. Ian Smith completed his PhD at Prince Henry's Institute Melbourne and Monash University, Australia. He is the vice-provost (research and research infrastructure) of Monash University. His research applies proteomics technologies to study the proteases involved in the generation and metabolism of peptide regulators involved in both brain and cardiovascular function.

Geoffrey I. Webb received his PhD degree in 1987 from La Trobe University. He is the director of the Monash Centre for Data Science and Professor in Faculty of Information Technology at Monash University, Australia. His research interests include machine learning, data mining, computational biology and user modelling.

Jian Li is a Professor and group leader in Monash Biomedicine Discovery Institute and Department of Microbiology, Monash University, Australia. He is a Web of Science 2015-17 Highly Cited Researcher in Pharmacology & Toxicology. He is currently an NHMRC Principal Research Fellow. His research interests include the pharmacology of polymyxins and the discovery of novel, safer polymyxins.

Jiangning Song is an Associate Professor and group leader in the Monash Biomedicine Discovery Institute and Biochemistry and Molecular Biology, Monash University, Australia. He is a member of the Monash Centre for Data Science, Monash University. His research interests include artificial intelligence, bioinformatics, machine learning, big data analytics, and pattern recognition.

Junfeng Xia is a professor in the Institutes of Physical Science and Information Technology and School of Computer Science and Technology, Anhui University. His research interests are bioinformatics and machine learning.