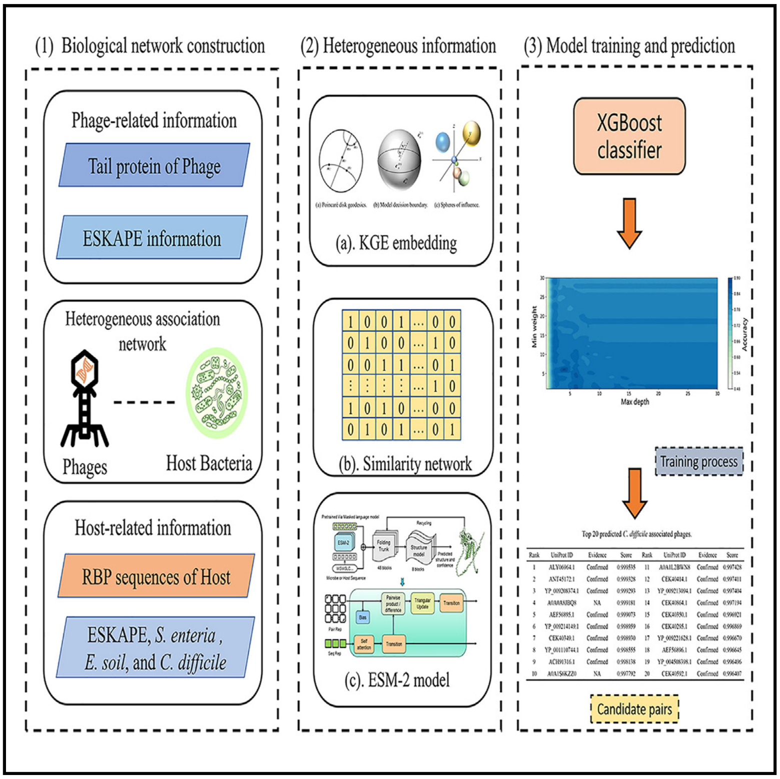
iScience

Article

Predicting phage-host interaction via hyperbolic Poincaré graph embedding and large-scale protein language technique

Graphical abstract



1	0	0	1	…	0	0
0	1	0	0	…	1	0
0	0	1	1	…	0	1
⋮	⋮	⋮	⋮	…	1	0
1	0	1	0	…	0	0
0	1	0	1	…	0	1

Top 20 predicted *C. difficile* associated phages.

Rank	UniProt ID	Evidence	Score	Rank	UniProt ID	Evidence	Score
1	ALY06964.1	Confirmed	0.999535	11	A0A1L2BWN8	Confirmed	0.997428
2	ANT45172.1	Confirmed	0.999328	12	CEK40414.1	Confirmed	0.997411
3	YP_009208374.1	Confirmed	0.999293	13	YP_009213094.1	Confirmed	0.997404
4	A0A0A8JBQ8	NA	0.999181	14	CEK40864.1	Confirmed	0.997194
5	AEF56895.1	Confirmed	0.999073	15	CEK40350.1	Confirmed	0.996921
6	YP_009214149.1	Confirmed	0.998959	16	CEK40295.1	Confirmed	0.996869
7	CEK40349.1	Confirmed	0.998930	17	YP_009221628.1	Confirmed	0.996670
8	YP_001110744.1	Confirmed	0.998555	18	AEF56896.1	Confirmed	0.996645
9	ACH91316.1	Confirmed	0.998138	19	YP_004508398.1	Confirmed	0.996496
10	A0A1S6KZZ0	NA	0.997792	20	CEK40592.1	Confirmed	0.996407

Highlights

- GE-PHI integrates knowledge graph embedding and language models to predict PHI
- The model incorporates microbial networks to capture complex relationships
- The prediction performance of GE-PHI demonstrated its robust performance
- Molecular docking further proves the robustness of our model

Authors

Jie Pan, Rui Wang, Wenjing Liu, ..., Jie Feng, Yanmei Sun, Shiwei Wang

Correspondence

sunyanmei@nwu.edu.cn (Y.S.),
wangsw@nwu.edu.cn (S.W.)

In brief

Virology; Microbiology; Bacteriology; Machine learning

Pan et al., 2025, iScience *28*, 111647
January 17, 2025 © 2024 Published by Elsevier Inc.
https://doi.org/10.1016/j.isci.2024.111647

iScience

CellPress
OPEN ACCESS

Article

Predicting phage-host interaction via hyperbolic Poincaré graph embedding and large-scale protein language technique

Jie Pan,[1,6] Rui Wang,[2,6] Wenjing Liu,[1] Li Wang,[1] Zhuhong You,[3] Yuechao Li,[3] Zhemeng Duan,[1] Qinghua Huang,[4] Jie Feng,[5] Yanmei Sun,[1,7,*] and Shiwei Wang[1,*]

[1]Key Laboratory of Resources Biology and Biotechnology in Western China, Ministry of Education, Provincial Key Laboratory of Biotechnology of Shaanxi Province, the College of Life Sciences, Northwest University, Xi'an 710069, China

[2]Department of Ophthalmology, The First Affiliated Hospital of Northwest University, 30 Fenxiang, the South Avenue, Xi'an, Shaanxi 710002, China

[3]School of Computer Science, Northwestern Polytechnical University, Xi'an 710129, China

[4]School of Artificial Intelligence, OPtics and ElectroNics (iOPEN), Northwestern Polytechnical University, Xi'an, Shaanxi 710072, China

[5]State Key Laboratory of Microbial Resources, Institute of Microbiology, Chinese Academy of Sciences, Beijing, China

[6]These authors contributed equally

[7]Lead contact

*Correspondence: sunyanmei@nwu.edu.cn (Y.S.), wangsw@nwu.edu.cn (S.W.)

https://doi.org/10.1016/j.isci.2024.111647

SUMMARY

Bacteriophages (phages) are increasingly viewed as a promising alternative for the treatment of antibiotic-resistant bacterial infections. However, the diversity of host ranges complicates the identification of target phages. Existing computational tools often fail to accurately identify phages across different bacterial species. In this study, we present GE-PHI, a machine-learning-based model for predicting phage-host interactions (PHIs) by integrating knowledge graph embedding algorithm with a large-scale protein language model. First, a phage-host heterogeneous association network (PHAN) was constructed that incorporated phage-phage and host-host similarity networks. Then, the multi-relational Poincaré graph embedding (MuRP) was used to extract topological patterns. Additionally, we employed the ESM-2 protein language model to capture evolutionary information from phage tail proteins and host-receptor-binding proteins. GE-PHI achieved a cross-validation area under the curve (AUC) of up to 0.9453 in silico and maintains this performance in case studies. This study provides insights into machine-learning-guided phage therapeutics and diagnostics in microbial engineering.

INTRODUCTION

Existing studies have suggested that bacterial infections are associated with a variety of diseases, such as cancer,[1] pneumonia,[2] asthma,[3] and as so on. Since the discovery of the first penicillin in 1928, various antibiotics have been extensively applied to treat bacterial infections in human and animals.[4] Unfortunately, many bacteria have developed drug-resistance mechanisms due to the extensive misuse of antibiotics, leading to the emergence of super bacteria.[5] The Centers for Disease Control and Prevention (CDC) in the United States reported an estimated 2.8 million cases of infections resistant to antibiotics occurring each year. According to the investigation in *The Lancet*, more than 1.27 million people died from antimicrobial resistance in 2019.[6] In addition, the European CDC reported that 35,000 people die each year from antibiotic-resistant infections.[7] Even worse, many pharmaceutical companies have ceased developing new antibiotics due to the high cost and time-consuming process. Therefore, it is crucial to develop alternative treatments to curb the continued spread of antibiotic-resistant infections.

Bacteriophages (phages) are among the most abundant and diverse organisms on Earth, which are capable of not only eliminating specific bacteria hosts but also replicating themselves.[8] Because of these properties, phage therapy has garnered increasing attention in recent years.[9] Identifying phage-host interactions (PHIs) can help researchers find candidate phages and determine their hosts ranges. However, most existing microbiological methods are intensive and cumbersome. As a result, scientists are devoted to provide computational methods for predicting PHIs, facilitating *in vivo* validation and contributing to the treatment of bacterial infections.[10]

The co-evolutionary processes at the molecular and ecological levels have shaped the genomes of phages and bacteria, leaving distinct signatures within their genomic sequences that researchers can utilize to predict PHIs.[11,12] Consequently, many computational approaches have been proposed to predict PHIs using genomic sequence information.[13] For example, Song et al. developed Prophage Hunter, which identifies active prophages from bacterial genomes by combining sequence-similarity-based matching with machine learning classification.[14]

© 2024 Published by Elsevier Inc.
This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

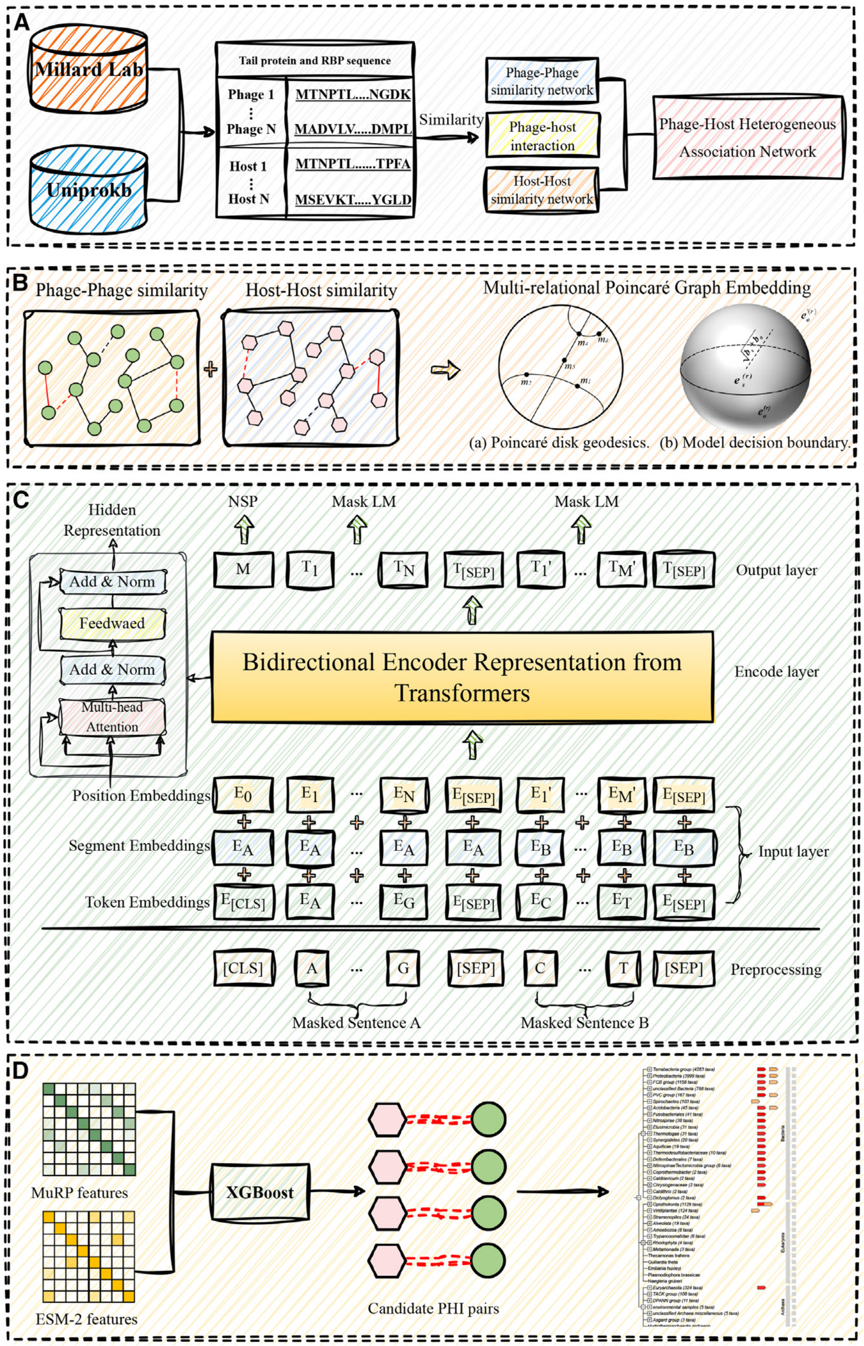


Figure 1. The flowchart of GE-PHI

(A) The construction of heterogeneous association network.

(B) The extraction process of knowledge-graph-embeddings-based topology structure information.

(C) BERT-based ESM-2 training process for encoding protein sequences information.

(D) The training and predicting process of XGBoost.

Boeckaerts et al. proposed a machine learning model called PhageHostLearn, which utilized sequence information to predict initial phage-*Klebsiella* pairs at the strain level.[15] Wang et al. proposed DeepHost model, which uses spaced k-mer pairs to account for sequence variations. This model is particularly effective for genomes with limited homology to existing datasets and offers faster processing than BLAST.[16] Despite these excellent performance in field of PHI prediction, several challenges remain.[17] First, these methods struggle to predict interactions involving novel or highly divergent phages.[18] Second, many methods lack the ability to integrate multi-modal data, such as protein-protein interactions (PPIs) or environmental context, which can provide more comprehensive insights into PHIs. Moreover, these methods lack interpretability.[19,20] For instance, deep-learning-based models operate as "black boxes" that make it difficult to explore underlying biological mechanisms.[21]

Recently, knowledge graph embedding (KGE) algorithms have attracted considerable attention in many bioinformatics fields, such as drug-drug interaction prediction,[22] transcription factor-gene prediction,[23] and polypharmacy side-effect prediction.[24] KGE can represent complex biological relationships in a low-dimensional space, concurrently retaining inherent structure and connectivity. This capability makes KGE particularly well suited for integrating diverse types of biological information (e.g., proteins, drugs, and genomes) into a unified predictive framework.[25] Additionally, KGE algorithms have the potential to capture both direct and indirect relationships between microbial entities, such as phages, bacterial hosts, and their associated genomic features.

Inspired by abovementioned excellent techniques, we propose a computational framework named GE-PHI, as shown in Figure 1, which predicts potential PHIs by integrating knowledge graph embedding with a large-scale protein language model. Specifically, we first construct a phage-host heterogeneous association network (PHAN) by incorporating phage-phage and host-host similarity networks into a unified interaction network. To capture the intricate microbial relationships within PHAN, we employed the multi-relational Poincaré graph embedding (MuRP) algorithm, a state-of-the-art KGE technique that leverages hyperbolic geometry to model the hierarchical and multi-relational structure of PHAN. In parallel, an advanced protein language model, ESM-2, was used to extract rich sequence features from phage tail proteins and

Table 1. Prediction performance of GE-PHI on the PHI task under 5-fold CV scheme

	Acc	F1	Pre	Rec	MCC	AUC	AUPR
Fold 1	0.8688	0.8724	0.8465	0.9000	0.7390	0.9446	0.9455
Fold 2	0.8974	0.9013	0.8725	0.9319	0.7965	0.9547	0.9548
Fold 3	0.8665	0.8682	0.8571	0.8796	0.7332	0.9301	0.9280
Fold 4	0.8953	0.8974	0.8794	0.9162	0.7913	0.9525	0.9518
Fold 5	0.8997	0.9026	0.8800	0.9263	0.8006	0.9447	0.9234
mean	0.8855	0.8884	0.8671	0.9108	0.7721	0.9453	0.9407
Std	0.0164	0.0167	0.0148	0.0212	0.0331	0.0086	0.0127

host-receptor-binding proteins (RBPs). The topological insights from MuRP and evolutionary information from ESM-2 are combined to create comprehensive feature vectors, which are subsequently integrated and processed by XGBoost classifier for final training and prediction. Experimental results demonstrated that our model outperforms state-of-the-art methods and some baseline algorithms. Ablation studies and case study experiments indicate the superior predictive accuracy and robustness of GE-PHI. This integrated approach not only enhances the precision of PHI predictions but also provides profound insights into the evolutionary and functional dynamics of PHIs, advancing the discovery of therapeutic phages and contributing significantly to global efforts against antibiotic resistance.

RESULTS

Prediction evaluation of the GE-PHI model

The performance of the proposed model was evaluated on the PHI dataset under 5-fold cross-validation strategy, and the detailed experimental results are summarized in Table 1. As visible from the table, GE-PHI achieved an average prediction accuracy of 0.8855. The accuracy of five experiments was 0.8688, 0.8974, 0.8665, 0.8953, and 0.8997, and the standard deviation was 0.0164. At the F1-score, Precision, Recall, and MCC scores, GE-PHI yielded 0.8884, 0.8671, 0.9108, and 0.7721, with standard deviations (std) of 0.0167, 0.0148, 0.0212, and 0.0331, respectively, with the highest AUC score reaching 0.9549. The ROC and area under the precision-recall (AUPR) curves are presented in Figure 2. These results not only demonstrate that the proposed GE-PHI model performs outstandingly on the phage-host prediction task but also suggest that it generalizes well across different folds, indicating its reliability and potential for broader application in PHI prediction.

Comparison with state-of-the-art models

To demonstrate the effectiveness of the GE-PHI model, we compared it with the following state-of-the-art models.

PHIAF[26] used a generative adversarial network (GAN) module to construct data augmentation and a sequence-based feature fusion technique to predict PHIs. The combination of GAN and sequence information aims to enhance the diversity and robustness of the training samples and improve the prediction ability.

GSPHI[27] is a deep learning framework that performed PHI prediction based on graph embedding method and natural language processing module. It leverages the structural information in graphs and the contextual information from sequences, making it particularly effective in understanding microbial mechanisms.

WIsH[28] identifies candidate hosts for phages by analyzing their genomic sequences. It trains a Markov model to calculate the interaction scores between phages and hosts.

PredPHI[29] is a deep learning model that predicts PHIs based on protein sequence information and a convolutional neural network (CNN).

As illustrated in Figure 3, GE-PHI outperforms these methods in terms of both AUC and AUPR values. Specifically, compared with the next best model, GSPHI, our model achieves an impressive enhancement of 2.45% in AUC and 4.47% in AUPR. We attribute these improvements to several factors in GE-PHI framework. First, the utilization of MuRP encodes the topological information of PHAN into a hyperbolic space, which is particularly suitable for extracting the intricate relationships in complex microbial networks. Additionally, the incorporation of the ESM-2 module allows our model to capture the structural feature of amino acid sequences, which can lead to more robust and accurate results.

Parameter sensitivity analysis

In the XGBoost classifier, the max depth *n* controls the model's complexity by limiting the depth of the tree, whereas the min weight *k* helps prevent overfitting by specifying the minimum sum of instance weights in a child node. To achieve the optimal performance of GE-PHI, we conducted a comprehensive grid search to find the best parameters. During this process, both *n* and *k* were ranged from 1 to 30, and the prediction task was repeated 900 times.

The contour plot (Figure 4) illustrates the relationship between max depth, min weight, and the resulting accuracy of the model. It can be observed that as the depth *n* increases, the accuracy improves until reaching a saturation point, beyond which the model's performance begins to stabilize. This trend indicates that deeper trees provide a more detailed representation of the data but can reach a point of diminishing returns. Similarly, for min weight *k*, increasing values initially lead to improvements in accuracy, likely due to the mitigation of overfitting by restricting the influence of low-weight nodes. However, after a certain threshold, further increases in *k* lead to a plateau, where additional constraints do not result in significant performance gains.

The plot highlights the optimal parameter combination, where $n = 4$ and $k = 6$, which yields the highest accuracy. This suggests that the model benefits from moderate tree complexity and a

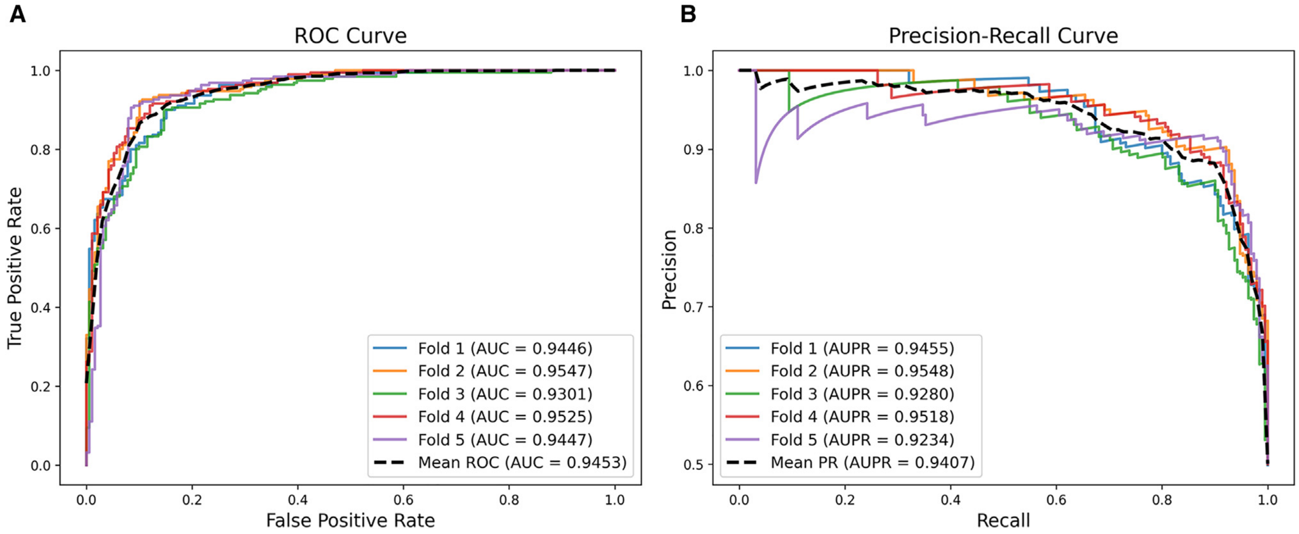


Figure 2. The ROC and AUPR curves obtained from the GE-PHI model on the PHI dataset

(A) The ROC curves obtained from GE-PHI model.

(B) The AUPR curves obtained from the GE-PHI model.

balanced restriction on node weight, leading to a well-regularized model that avoids overfitting while capturing the essential patterns in the PHI data. The smooth upward slope in accuracy as *n* and *k* approach their optimal values confirms that the GE-PHI model is robust to small changes in these parameters, providing reliable performance across a range of settings.

Prediction evaluation on independent dataset

To assess the generalization ability of GE-PHI, we applied it to an independent dataset that is constructed by *Acinetobacter baumannii*. Specifically, we excluded all data related to *A. baumannii* and its phages from the original dataset, using them exclusively as the test set. The remaining data were used to train the prediction model. In this way, the overlaps in hosts and phages in the training and testing sample can be avoided. This approach also evaluates the model's ability to predict interactions involving unseen bacteria or phages and tests the robustness and applicability to scenarios.

The detailed experimental results are presented in Table 2. As shown in the table, GE-PHI yielded average Acc, AUC, and AUPR values of 0.7951, 0.8904, and 0.8874, respectively. Although these results are slightly lower than the performance achieved on the original dataset, the decrease is expected due to the inherent variability and complexity of independent test data. The independent dataset contains unique features that were not present during model training, thus posing a greater challenge. Despite this, the model still performs well, demonstrating its robustness. These findings suggest that GE-PHI can effectively predict PHIs, even for untraining biological information.

Comparison of MuRP with some baseline algorithms

In our study, to further validate the effectiveness of the MuRP algorithm, we compared it with some popular knowledge graph embedding methods, including RotatE,[30] TransE,[31] TransF,[32] DistMult,[33] ComplEx,[34] HolE,[35] and SimplE.[36] All baseline algorithms were implemented with their default parameters, as recommended in their original publications, which can ensure consistency and unbiased performance evaluation across the different methods.

As shown in Figure 5, the experimental results indicate that GE-PHI achieves the best prediction performance. Furthermore, GE-PHI yielded a small standard deviation for all the baseline algorithms, which demonstrates that MuRP is not only accurate but also reliable across different data splits. When focusing on specific metrics, GE-PHI exhibited significant improvements in Rec and MCC metrics. Compared with the best results among the seven baseline algorithms, Rec and MCC improved by 7.03% (84.05%–91.08%) and 10.59% (66.62%–77.21%), respectively. The detailed 5-fold cross validation (CV) results of these KGE-based algorithms are listed in supplemental materials Tables S1–S7. These increasements are substantial and highlight the efficacy of MuRP in capturing the complex relational patterns in heterogeneous networks. Moreover, the utilization of the MuRP algorithm within the GE-PHI model enabled the model to maintain the high performance while minimizing the variance across different CV folds.

Comparison of XGBoost with other machine learning classifiers

In the proposed GE-PHI model, XGBoost was adopted as the classifier for final training and prediction. To validate the effectiveness of XGBoost in predicting candidate PHI pairs, we conducted a comprehensive comparison with five traditional machine learning classifiers, including support vector machine (SVM), k-nearest neighbor (KNN), Decision Tree (DT), Random Forest (RF), and gradient boosting decision trees (GBDT). The results of this comparison are summarized in Figure 6, and the detailed 5-fold CV results of different classifiers are listed in the Tables S8–S12.

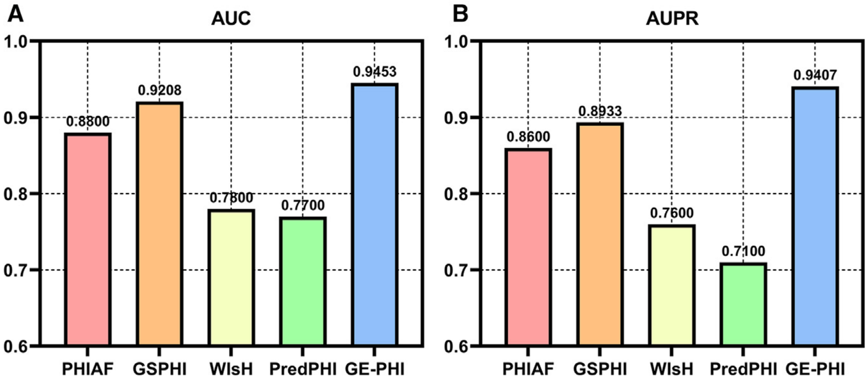


Figure 3. Comparison of the predictive performance of our proposed GE-PHI model and other state-of-the-art methods on the PHI dataset in terms of AUC and AUPR value

(A) The experimental results of AUC values.

(B) The experimental results of AUPR values.

As shown in the abovementioned figure, it can be observed that XGBoost significantly outperformed all other machine learning classifiers. For instance, compared with the best performing classifier, RF, the GE-PHI model achieved a notable improvement in accuracy by 3.59% (0.8496–0.8855). Moreover, it also yielded superior performance across six other evaluation metrics, which demonstrated its effectiveness in the PHI prediction task. The excellent performance of XGBoost can be attributed to several factors. First, XGBoost is a gradient boosting model that each subsequent model attempts to correct the errors of the previous ones. This iterative strategy improves accuracy and helps avoid overfitting, which is particularly beneficial for complex bioinformatics tasks. Second, XGBoost incorporates a regularization technique to control the model complexity. Compared with traditional machine learning classifiers (e.g., DT or RF, even GBDT), this regularization capability is crucial in bioinformatics for dealing with noisy high-dimensional microbial data. Compared with traditional machine learning classifiers, we have strong reasons to believe that the choice of XGBoost as the classifier in the GE-PHI model is highly feasible.

Ablation studies

In this section, an ablation study is conducted to investigate the contribution of two main components in the GE-PHI model, including the ESM-2 module for embedding protein evolutionary information and the MuRP model for capturing the topological information from microbial interactions network. Two variants of the GE-PHI model were constructed, and the comparison results are shown in Table 3 and Figure 7.

GE-ESM: a variant of the GE-PHI model that only used the protein sequence information for the PHI prediction task.

GE-MuRP: a variant model of GE-PHI that only used the topological structure information for the PHI prediction task.

From Table 3, the following observations can be obtained. First, the variant GE-MuRP model performs worse than GE-ESM, indicating that although the topology information of PHAN is valuable, the evolutionary features from protein sequences are even more crucial. This phenomenon highlights the importance of protein evolutionary context in understanding the molecular mechanism. Second, both variant models perform worse than GE-PHI, which suggests that the proposed method is reasonable and effective. The superior performance

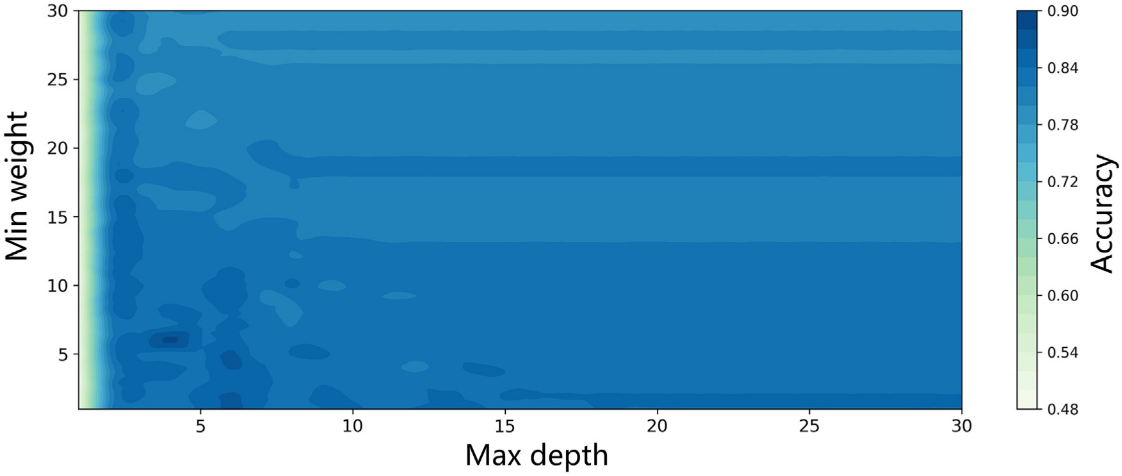


Figure 4. Accuracy values formed by the parameter max depth *n* and min weight *k*

Table 2. Prediction performance of GE-PHI on the independent *A. baumannii* species

	Acc	F1	Pre	Rec	MCC	AUC	AUPR
Fold 1	0.8294	0.8303	0.8238	0.8368	0.6589	0.9091	0.9061
Fold 2	0.7711	0.7740	0.7680	0.7801	0.5421	0.8789	0.8815
Fold 3	0.7958	0.7903	0.8122	0.7696	0.5924	0.892	0.8834
Fold 4	0.8063	0.8103	0.794	0.8272	0.6131	0.9106	0.9157
Fold 5	0.7731	0.7749	0.7708	0.7789	0.5462	0.8612	0.8504
mean	0.7951	0.7960	0.7938	0.7985	0.5905	0.8904	0.8874
Std	0.0243	0.0242	0.0247	0.0310	0.0487	0.0209	0.0253

of GE-PHI further indicated that both the protein sequence information embedded by ESM-2 and topology structure information captured by MuRP contribute significantly to the GE-PHI model. It also reveals that the sequences information plays a more important role in these specific bioinformatics tasks. These observations demonstrated the necessity of incorporating both types of information.

Case studies

To further assess the practical effectiveness of our model in predicting associations between phages and hosts, we performed case studies on both gram-positive and gram-negative bacteria. For a given host, we excluded all known phage interactions from the dataset. The proposed model was then trained on the remaining known interaction pairs and tested on candidate samples to identify phages associated with the target bacteria. This approach ensures independence between the training and validation sets. In other words, the prediction model relies solely on the remaining interaction pairs and the information captured during training and testing. Specifically, in the phage-host adjacency matrix A, the known interactions for the selected hosts were changed from 1 to 0, and the candidate phages were ranked based on their prediction scores. The top 20 ranked phages were further validated using relevant databases. The model is considered effective if the top-ranked predictions include a higher number of verified interactions. To investigate the complex interactions between phages and hosts, we conducted independent case studies on two significant bacterial hosts: *Clostridium difficile* and *Escherichia coli*.

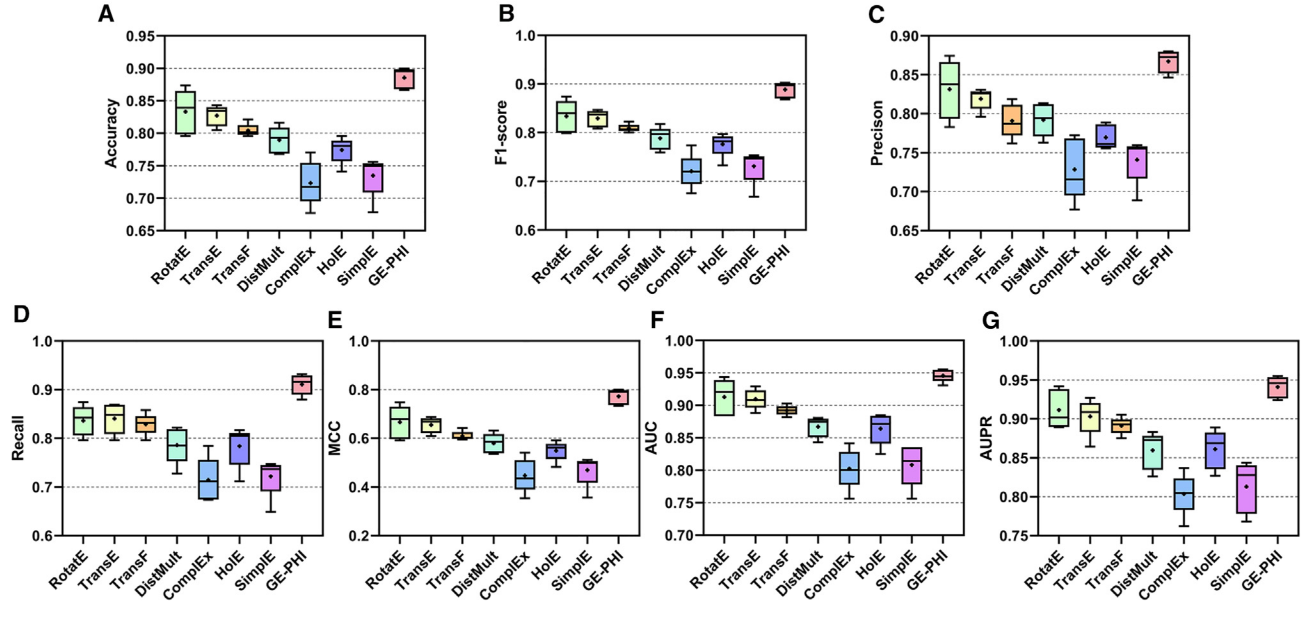


Figure 5. The comparison results of MuRP with some knowledge graph embeddings methods

(A) The comparison result of accuracy.
(B) The comparison result of F1-score.
(C) The comparison result of precision.
(D) The comparison result of recall.
(E) The comparison result of MCC.
(F) The comparison result of AUC values.
(G) The comparison result of AUPR values.

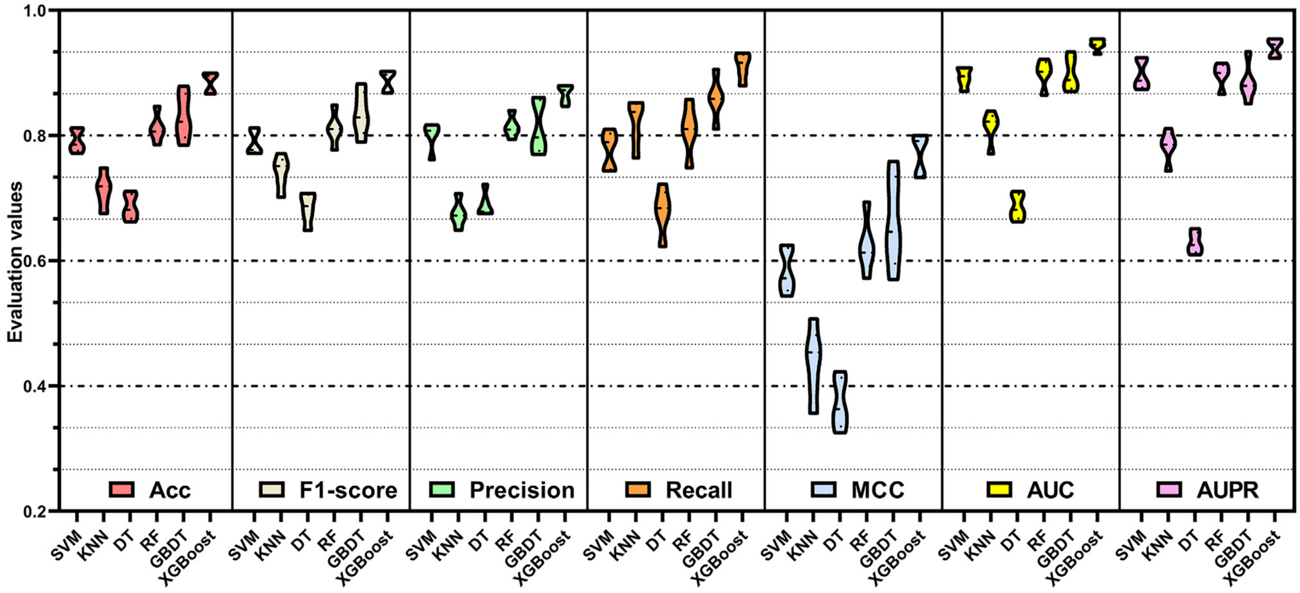


Figure 6. The comparison results of accuracy, F1-score, precision, recall, MCC, AUC, and AUPR values on PHI prediction task

C. difficile is gram-positive bacteria, which is the primary pathogen responsible for antibiotic-associated diarrhea and pseudomembranous colitis. The infection symptoms included vomiting, abdominal pain, and fever, and in severe cases, it can lead to dehydration, intestinal perforation, sepsis, and even death. In this study, the top 20 prediction results of *C. difficile*-interacted phages are listed in Table 4. We can see that 18 of the top 20 predicted samples have been verified by public database. *E. coli* is a kind of gram-negative bacteria that is the predominant and most abundant bacteria in the intestinal tract of human and animals. In the laboratory of microbiological research, *E. coli* is one of the main model organisms. Similar to *C. difficile*, the top 20 predicted results of *E. coli*-interacted phages are summarized in Table 5. Based on the public databases, 17 of the top 20 predicted phages could be verified that they are related to *E. coli*. These case study results demonstrate that GE-PHI can help to identify PHI pairs and narrow the scope of candidates for further biological experiments.

Molecular docking experiments

To further validate the reliability of our model, we conducted molecular docking experiments on the unconfirmed phages listed in Tables 4 and 5. These experiments were performed to evaluate the interaction strength between the predicted phages and their corresponding host receptors. First, protein sequences were input into the Alphafold 3[37] server (https://alphafoldserver.com/) to generate the corresponding protein structure files in PDB format. Next, molecular docking was performed using the GRAMM web server (https://gramm.compbio.ku.edu/gramm), and the docking results were analyzed and visualized through PyMOL software. To assess the strength of the interactions, binding energy was selected as the evaluation metric, where lower binding energy indicates stronger molecular interactions and more reliable predictions.

For *C. difficile*-associated phages shown in Table 4, two unconfirmed phages, *A0A0A8JBQ8* and *A0A1S6KZZ0*, were selected for docking analysis. As shown in Figure 8, *A0A0A8JBQ8* exhibited a binding energy of −8.1 kcal/mol with its host receptor, whereas *A0A1S6KZZ0* had a stronger binding energy of −11.1 kcal/mol. These results suggest that these phages have significant potential to interact with *C. difficile*. For *E. coli*-associated phages shown in Table 5, *A0A346FJ10* exhibited a binding energy of −16.6 kcal/mol, whereas *A0A0H4TFM9* and *YP_009201899.1* showed stronger binding energies of −35.2 kcal/mol and −32.6 kcal/mol, respectively. These values highlight the strong interaction potential of these phages with *E. coli*, further validating the model's predictions.

The docking results provide important insights into the molecular interactions between the predicted phages and their target hosts. The low binding energies observed for these unconfirmed

Table 3. Ablation studies of GE-PHI model with its two variant model

Model	Acc	F1	Pre	Rec	MCC	AUC	AUPR
GE-ESM	0.8593 ± 0.0154	0.8586 ± 0.0178	0.8624 ± 0.0094	0.8552 ± 0.0317	0.7189 ± 0.0306	0.9381 ± 0.017	0.9352 ± 0.0235
GE-MuRP	0.8094 ± 0.0149	0.818 ± 0.0153	0.7833 ± 0.0147	0.8563 ± 0.0244	0.6217 ± 0.0301	0.8785 ± 0.0109	0.8459 ± 0.0211
GE-PHI	0.8855 ± 0.0164	0.8884 ± 0.0167	0.8671 ± 0.0148	0.9108 ± 0.0212	0.7721 ± 0.0331	0.9458 ± 0.0098	0.9410 ± 0.0140

Data are represented as mean ± SD.

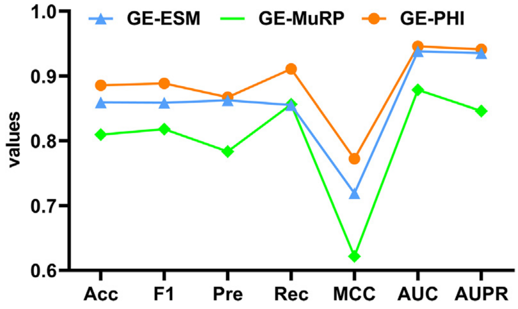


Figure 7. The comparison results of GE-PHI with its two variant models

phages suggest strong binding affinity, reinforcing the reliability of the GE-PHI model in identifying PHIs and providing a strong foundation for further experimental investigations. Moreover, these findings emphasize the importance of integrating computational predictions with molecular validation, paving the way for the development of phage-based therapies targeting antibiotic-resistant bacteria.

DISCUSSION

In this work, we developed GE-PHI, a comprehensive computational model to address several key challenges in PHI prediction, particularly in the context of phage therapeutics and diagnostics. GE-PHI effectively predicts PHIs by integrating advanced knowledge graph embeddings and large-scale protein language models. The model generates robust prediction scores, which can be used to prioritize candidates for laboratory validation. Additionally, case studies on both gram-positive and gram-negative bacteria further demonstrated the applicability and reliability of our model in diverse real-world scenarios.

The construction of PHAN in GE-PHI presents notable strengths. One of the primary advantages of PHAN is its ability to integrate multiple types of relationships, including phage-phage and host-host similarities into a unified microbial graph. This comprehensive network structure enables the model to capture complex interactions that might be overlooked. Moreover, the similarity networks play a crucial role in enhancing the predictive power. These similarity networks allow GE-PHI to identify potential PHIs by leveraging shared features and patterns across different phages and hosts, which can enrich additional relational information.

A powerful KGE algorithm, MuRP, was used to capture the topological information from PHAN. It is highly effective at modeling hierarchical and multi-relational data, which is particularly suitable for the complex structure of PHAN. Compared with traditional Euclidean-based methods, MuRP can be used more accurately to preserve the relationships between different entities. To be more specific, when compared with common knowledge graph embedding techniques such as TransE, RotatE, and ComplEx, MuRP demonstrates a superior ability to capture complex network structures, primarily due to the application of hyperbolic space.

The ESM-2 algorithm utilizes large-scale evolutionary data to generate high-dimensional embeddings that encapsulate both the topological structure and functional features from phage tail protein and host-receptor-binding proteins. Compared with traditional methods, ESM-2 can extract a more nuanced and comprehensive representation of protein sequences. When the sequence homology is low, the ability of ESM-2 to capture subtle sequence variations is particularly valuable.

In summary, the proposed GE-PHI model represents a significant advancement in the field of PHI prediction. The construction of the PHAN network allows for a comprehensive analysis of both direct and indirect relationships between different microbial nodes. The utilization of MuRP provides a robust way for capturing hierarchical and multi-relational data, whereas ESM-2 contributes detailed sequence insights that enhance the model’s ability to predict interactions even in cases of low sequence homology. Additionally, the integration of these features into an XGBoost classifier ensures that our model is not only accurate but also efficient, leveraging the strengths of gradient boosting to handle complex information. Despite these strengths, there are still some challenges in optimizing and interpreting the model, particularly in the computational demands and the black-box nature of deep learning components. Nevertheless, case studies and molecular docking results in both gram-positive and gram-negative bacteria further demonstrated its potential in predicting unknown PHI pairs. We hope that our model can be a valuable tool for advancing phage therapy and reducing antibiotic resistance.

Table 4. Top 20 predicted *C. difficile*-associated phages

Rank	UniProt ID	Evidence	Score	Rank	UniProt ID	Evidence	Score
1	ALY06964.1	Confirmed	0.999535	11	A0A1L2BWN8	Confirmed	0.997428
2	ANT45172.1	Confirmed	0.999328	12	CEK40414.1	Confirmed	0.997411
3	YP_009208374.1	Confirmed	0.999293	13	YP_009213094.1	Confirmed	0.997404
4	A0A0A8JBQ8	NA	0.999181	14	CEK40864.1	Confirmed	0.997194
5	AEF56895.1	Confirmed	0.999073	15	CEK40350.1	Confirmed	0.996921
6	YP_009214149.1	Confirmed	0.998959	16	CEK40295.1	Confirmed	0.996869
7	CEK40349.1	Confirmed	0.998930	17	YP_009221628.1	Confirmed	0.996670
8	YP_001110744.1	Confirmed	0.998555	18	AEF56896.1	Confirmed	0.996645
9	ACH91316.1	Confirmed	0.998138	19	YP_004508398.1	Confirmed	0.996496
10	A0A1S6KZZ0	NA	0.997792	20	CEK40592.1	Confirmed	0.996407

Table 5. Top 20 predicted *E. coli*-associated phages

Rank	UniProt ID	Evidence	Score	Rank	UniProt ID	Evidence	Score
1	A0A0U4IIL0	Confirmed	0.999551	11	A0A2P1CL14	Confirmed	0.996766
2	A0A386K7G1	Confirmed	0.999372	12	Q7Y2W1	Confirmed	0.996537
3	A0A097J8C7	Confirmed	0.998553	13	A0A2P1CL03	Confirmed	0.996417
4	A0A097J306	Confirmed	0.998440	14	Q8LTV0	Confirmed	0.996184
5	D9IET6	Confirmed	0.998329	15	A0A0H4TFM9	NA	0.995637
6	A0A3G3MCX8	Confirmed	0.998183	16	A0A0A7HBH2	Confirmed	0.995250
7	A0A346FJ10	NA	0.997971	17	YP_009201899.1	NA	0.995226
8	A0A0M7QBC7	Confirmed	0.997424	18	A0A076G6W3	Confirmed	0.994862
9	P49714	Confirmed	0.997396	19	A0A0N7C1B3	Confirmed	0.994809
10	A0A0U4IB10	Confirmed	0.997346	20	A0A343S154	Confirmed	0.994759

Limitations of the study

Although the GE-PHI model demonstrates strong predictive capabilities, it has certain limitations. The reliance on cosine similarity in similarity networks may restrict the detection of non-linear relationships, which are crucial in some biological processes. Additionally, the model's complexity leads to high computational costs, particularly when processing large-scale datasets, which could limit its scalability. Although simpler Euclidean-based embedding methods, such as TransE or DistMult, are more computationally efficient, they did not perform as effectively as GE-PHI on high-dimensional or complex datasets. To address these challenges, future work will focus on improving the computational efficiency of GE-PHI through distributed computing and algorithm optimization. Furthermore, we plan to integrate additional biological data sources, such as environmental context and host gene expression data, to provide richer information and potentially enhance the model's predictive accuracy and robustness.

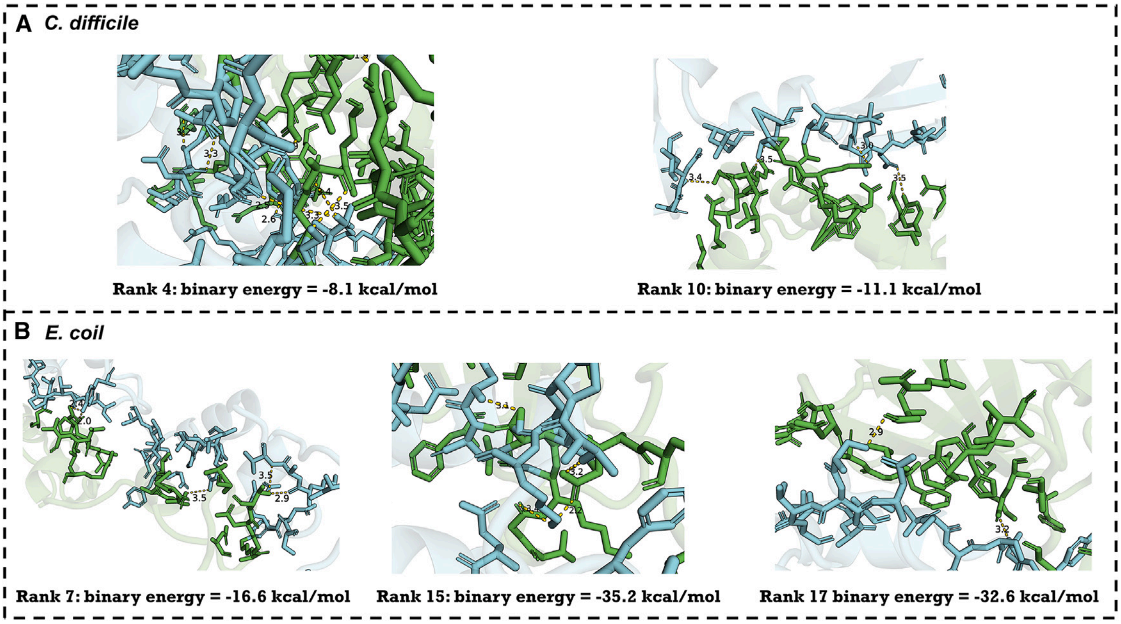
RESOURCE AVAILABILITY

Lead contact

The datasets and source code are available: https://github.com/JIENWU/GE-PHI. Inquires can be directed to the lead contact, Yanmei Sun (sunyanmei@nwu.edu.cn).

MATERIALS AVAILABILITY

All materials reported in this paper will be shared by the lead contact upon request.



C. difficile

Rank 4: binary energy = -8.1 kcal/mol

Rank 10: binary energy = -11.1 kcal/mol

E. coil

Rank 7: binary energy = -16.6 kcal/mol

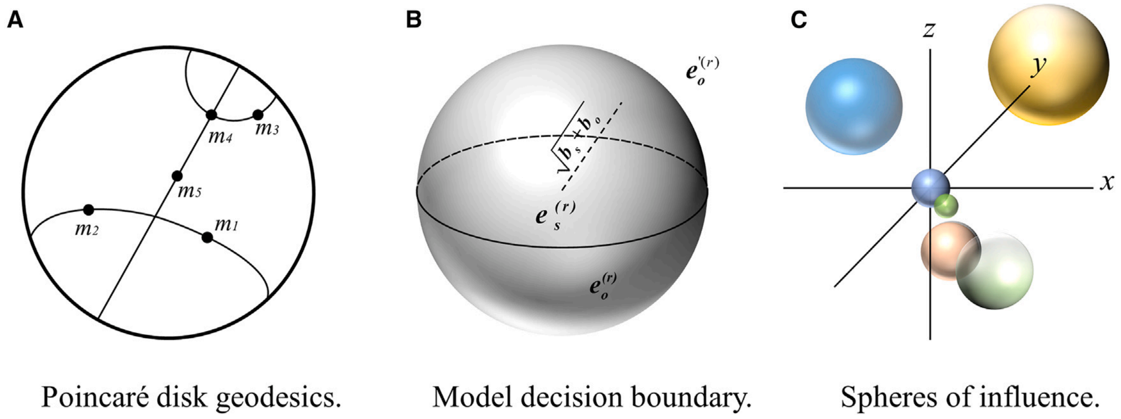
Rank 15: binary energy = -35.2 kcal/mol

Rank 17 binary energy = -32.6 kcal/mol

Figure 8. Molecular docking results of the candidate phages for gram-positive and gram-negative host

(A) The molecular docking results for *C. difficile*.

(B) The molecular docking results for *E. coli*.



Poincaré disk geodesics. Model decision boundary. Spheres of influence.

Figure 9. The architecture of MuRP algorithm

(A) Geodesics in the Poincaré disk that demonstrated the shortest paths between phages and bacteria nodes in PHAN.

(B) The model identifies the triple (e_s, r, e_o) as true and (e_s, r, e_o') as false.

(C) Each entity embedding has a sphere of influence, whose radius is determined by the entity-specific bias.

Data and code available

- Data reported in this paper will be shared by the lead contact upon request. This paper analyzes existing, publicly available data. These accession numbers for the datasets are listed in the key resources table.
- The datasets and source code are available: https://github.com/JIENWU/GE-PHI.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

ACKNOWLEDGMENTS

This work was supported in part by the grants from the Science & Technology Fundamental Resources Investigation Program (grant No. 2022FY101100) and the Foundation of Shaanxi Provincial Science and Technology Innovation Team (2024RS-CXTD-11). This work was also supported by the National Natural Science Foundation of China (grant No. 32170114 and 31770152) and the National Science Fund for Distinguished Young Scholars of China (grant No. 62325308). This work was also supported by Shaanxi Fundation of China Research Project for Chemistry & Biology (Grant No. 22JHZ008).

AUTHOR CONTRIBUTIONS

J.P., conceptualization, writing—original draft. R.W., Q.H., and Z.Y., methodology, supervision. W.L. and J.F., formal analysis. Z.D., L.W., and S.W., methodology, writing—review & editing. Y.L. and Y.S., visualization, writing—review & editing.

DECLARATION OF INTERESTS

The authors declare no competing interests.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- METHOD DETAILS
  - Phage-host heterogeneous association network
  - Hyperbolic graph embeddings for PHAN
  - Evolutionary scale modeling of sequence information
  - Machine learning-based classifier
  - Evaluation metrics
- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2024.111647.

Received: September 24, 2024
Revised: October 24, 2024
Accepted: December 17, 2024
Published: December 19, 2024

REFERENCES

1. El Tekle, G., and Garrett, W.S. (2023). Bacteria in cancer initiation, promotion and progression. Nat. Rev. Cancer *23*, 600–618. https://doi.org/10.1038/s41568-023-00594-2.
2. Gautam, S., Cohen, A.J., Stahl, Y., Valda Toro, P., Young, G.M., Datta, R., Yan, X., Ristic, N.T., Bermejo, S.D., Sharma, L., et al. (2020). Severe respiratory viral infection induces procalcitonin in the absence of bacterial pneumonia. Thorax *75*, 974–981. https://doi.org/10.1136/thoraxjnl-2020-214896.
3. Kraft, M. (2000). The role of bacterial infections in asthma. Clin. Chest Med. *21*, 301–313. https://doi.org/10.1016/S0272-5231(05)70268-9.
4. Hutchings, M.I., Truman, A.W., and Wilkinson, B. (2019). Antibiotics: past, present and future. Curr. Opin. Microbiol. *51*, 72–80. https://doi.org/10.1016/j.mib.2019.10.008.
5. Dong, X., Shi, P., Liu, W., Bai, J., and Bian, L. (2022). Metallo-beta-lactamase CphA evolving into more efficient hydrolases through gene mutation is a novel pathway for the resistance of super bacteria. Appl. Microbiol. Biotechnol. *106*, 2471–2480. https://doi.org/10.1007/s00253-022-11879-1.
6. Murray, C.J., Ikuta, K.S., Sharara, F., Swetschinski, L., Aguilar, G.R., Gray, A., Han, C., Bisignano, C., Rao, P., and Wool, E. (2022). Global burden of bacterial antimicrobial resistance in 2019: a systematic analysis. Lancet (North Am. Ed.) *399*, 629–655. https://doi.org/10.1016/S0140-6736(21)02724-0.

7. Cassini, A., Högberg, L.D., Plachouras, D., Quattrocchi, A., Hoxha, A., Simonsen, G.S., Colomb-Cotinat, M., Kretzschmar, M.E., Devleesschauwer, B., Cecchini, M., et al. (2019). 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. *19*, 56–66. https://doi.org/10.1016/S1473-3099(18)30605-4.

8. Dion, M.B., Oechslin, F., and Moineau, S. (2020). Phage diversity, genomics and phylogeny. Nat. Rev. Microbiol. *18*, 125–138. https://doi.org/10.1038/s41579-019-0311-5.

9. Suh, G.A., Lodise, T.P., Tamma, P.D., Knisely, J.M., Alexander, J., Aslam, S., Barton, K.D., Bizzell, E., Totten, K.M.C., Campbell, J.L., et al. (2022). Considerations for the use of phage therapy in clinical practice. Antimicrob. Agents Chemother. *66*, e0207121. https://doi.org/10.1128/aac.02071-21.

10. Edwards, R.A., McNair, K., Faust, K., Raes, J., and Dutilh, B.E. (2016). Computational approaches to predict bacteriophage–host relationships. FEMS Microbiol. Rev. *40*, 258–272. https://doi.org/10.1093/femsre/fuv048.

11. Safari, F., Sharifi, M., Farajnia, S., Akbari, B., Karimi Baba Ahmadi, M., Negahdaripour, M., and Ghasemi, Y. (2020). The interaction of phages and bacteria: the co-evolutionary arms race. Crit. Rev. Biotechnol. *40*, 119–137. https://doi.org/10.1080/07388551.2019.1674774.

12. Pan, J., Zhang, Z., Li, Y., Yu, J., You, Z., Li, C., Wang, S., Zhu, M., Ren, F., Zhang, X., et al. (2024). A microbial knowledge graph-based deep learning model for predicting candidate microbes for target hosts. Briefings Bioinf. *25*, bbae119. https://doi.org/10.1093/bib/bbae119.

13. Nie, W., Qiu, T., Wei, Y., Ding, H., Guo, Z., and Qiu, J. (2024). Advances in phage–host interaction prediction: in silico method enhances the development of phage therapies. Briefings Bioinf. *25*, bbae117. https://doi.org/10.1093/bib/bbae117.

14. Song, W., Sun, H.-X., Zhang, C., Cheng, L., Peng, Y., Deng, Z., Wang, D., Wang, Y., Hu, M., Liu, W., et al. (2019). Prophage Hunter: an integrative hunting tool for active prophages. Nucleic Acids Res. *47*, W74–W80. https://doi.org/10.1093/nar/gkz380.

15. Boeckaerts, D., Stock, M., Ferriol-González, C., Oteo-Iglesias, J., Sanjuán, R., Domingo-Calap, P., De Baets, B., and Briers, Y. (2024). Prediction of Klebsiella phage-host specificity at the strain level. Nat. Commun. *15*, 4355. https://doi.org/10.1038/s41467-024-48675-6.

16. Ruohan, W., Xianglilan, Z., Jianping, W., and Shuai Cheng, L.I. (2022). DeepHost: phage host prediction with convolutional neural network. Briefings Bioinf. *23*, bbab385. https://doi.org/10.1093/bib/bbab385.

17. Coclet, C., and Roux, S. (2021). Global overview and major challenges of host prediction methods for uncultivated phages. Curr. Opin. Virol. *49*, 117–126. https://doi.org/10.1016/j.coviro.2021.05.003.

18. Araújo, P.H.M.A.M. (2021). Bacteriophage-host determinants: identification of bacteriophage receptors through machine learning techniques.

19. Li, J., Liu, P., Chen, L., Pedrycz, W., and Ding, W. (2024). An Integrated Fusion Framework for Ensemble Learning Leveraging Gradient Boosting and Fuzzy Rule-Based Models. IEEE Trans. Artif. Intell. *5*, 5771–5785. https://doi.org/10.1109/TAI.2024.3424427.

20. Li, H., Wang, C., and Huang, Q. (2024). Employing Iterative Feature Selection in Fuzzy Rule-Based Binary Classification. IEEE Trans. Fuzzy Syst. *32*, 5109–5121. https://doi.org/10.1109/TFUZZ.2024.3414836.

21. Gabel, J., Desaphy, J., and Rognan, D. (2014). Beware of Machine Learning-Based Scoring Functions; On the Danger of Developing Black Boxes. J. Chem. Inf. Model. *54*, 2807–2815. https://doi.org/10.1021/ci500406k.

22. Su, X., You, Z.H., Huang, D.s., Wang, L., Wong, L., Ji, B., and Zhao, B. (2022). Biomedical knowledge graph embedding with capsule network for multi-label drug-drug interaction prediction. IEEE Trans. Knowl. Data Eng. *35*, 5640–5651. https://doi.org/10.1109/TKDE.2022.3154792.

23. Wu, Y.-H., Huang, Y.-A., Li, J.-Q., You, Z.-H., Hu, P.-W., Hu, L., Leung, V.C.M., and Du, Z.-H. (2023). Knowledge graph embedding for profiling the interaction between transcription factors and their target genes. PLoS Comput. Biol. *19*, e1011207. https://doi.org/10.1371/journal.pcbi.1011207s.

24. Nováček, V., and Mohamed, S.K. (2020). Predicting polypharmacy side-effects using knowledge graph embeddings. AMIA Summits on Translational Science Proceedings *2020*, 449.

25. Alam, F., Giglou, H.B., and Malik, K.M. (2023). Automated clinical knowledge graph generation framework for evidence based medicine. Expert Syst. Appl. *233*, 120964. https://doi.org/10.1016/j.eswa.2023.120964.

26. Li, M., and Zhang, W. (2022). PHIAF: prediction of phage-host interactions with GAN-based data augmentation and sequence-based feature fusion. Briefings Bioinf. *23*, bbab348. https://doi.org/10.1093/bib/bbab348.

27. Pan, J., You, W., Lu, X., Wang, S., You, Z., and Sun, Y. (2023). GSPHI: a novel deep learning model for predicting phage-host interactions via multiple biological information. Comput. Struct. Biotechnol. J. *21*, 3404–3413. https://doi.org/10.1016/j.csbj.2023.06.014.

28. Galiez, C., Siebert, M., Enault, F., Vincent, J., and Söding, J. (2017). WIsH: who is the host? Predicting prokaryotic hosts from metagenomic phage contigs. Bioinformatics *33*, 3113–3114. https://doi.org/10.1093/bioinformatics/btx383.

29. Li, M., Wang, Y., Li, F., Zhao, Y., Liu, M., Zhang, S., Bin, Y., Smith, A.I., Webb, G.I., Li, J., et al. (2021). A deep learning-based method for identification of bacteriophage-host interaction. IEEE ACM Trans. Comput. Biol. Bioinf *18*, 1801–1810. https://doi.org/10.1109/TCBB.2020.3017386.

30. Sun, Z., Deng, Z.-H., Nie, J.-Y., and Tang, J. (2019). Rotate: Knowledge graph embedding by relational rotation in complex space. preprint at arXiv. https://doi.org/10.48550/arXiv.1902.10197.

31. Bordes, A., Usunier, N., Garcia-Duran, A., Weston, J., and Yakhnenko, O. (2013). Translating embeddings for modeling multi-relational data. Adv. Neural Inf. Process. Syst. *26*.

32. Feng, J., Huang, M., Wang, M., Zhou, M., Hao, Y., and Zhu, X. (2016). Knowledge graph embedding by flexible translation. In Fifteenth International Conference on the Principles of Knowledge Representation and Reasoning (AAAI Press).

33. Yang, B., Yih, W.-t., He, X., Gao, J., and Deng, L. (2014). Embedding entities and relations for learning and inference in knowledge bases. preprint at arXiv. https://doi.org/10.48550/arXiv.1412.6575.

34. Trouillon, T., Welbl, J., Riedel, S., Gaussier, É., and Bouchard, G. (2016). Complex embeddings for simple link prediction. In Proceedings of The 33rd International Conference on Machine Learning, pp. 2071–2080. https://doi.org/10.48550/arXiv.1606.06357.

35. Nickel, M., Rosasco, L., and Poggio, T. (2016). Holographic embeddings of knowledge graphs. Proceedings of the AAAI conference on artificial intelligence 30. https://doi.org/10.1609/aaai.v30i1.10314.

36. Kazemi, S.M., and Poole, D. (2018). Simple embedding for link prediction in knowledge graphs. Adv. Neural Inf. Process. Syst. *31*.

37. Abramson, J., Adler, J., Dunger, J., Evans, R., Green, T., Pritzel, A., Ronneberger, O., Willmore, L., Ballard, A.J., Bambrick, J., et al. (2024). Accurate structure prediction of biomolecular interactions with AlphaFold 3. Nature *630*, 493–500. https://doi.org/10.1038/s41586-024-07487-w.

38. Cook, R., Brown, N., Redgwell, T., Rihtman, B., Barnes, M., Clokie, M., Stekel, D.J., Hobman, J., Jones, M.A., and Millard, A. (2021). INfrastructure for a PHAge REference database: identification of large-scale biases in the current collection of cultured phage genomes. Phage *2*, 214–223. https://doi.org/10.1089/phage.2021.0007.

39. UniProt Consortium (2019). UniProt: a worldwide hub of protein knowledge. Nucleic Acids Res. *47*, D506–D515. https://doi.org/10.1093/nar/gky1049.

40. Yuan, G., Zhai, Y., Tang, J., and Zhou, X. (2023). CSCIM_FS: Cosine similarity coefficient and information measurement criterion-based feature selection method for high-dimensional data. Neurocomputing *552*, 126564. https://doi.org/10.1016/j.neucom.2023.126564.

41. Ghazi, A.R., Münch, P.C., Chen, D., Jensen, J., and Huttenhower, C. (2022). Strain identification and quantitative analysis in microbial communities. J. Mol. Biol. *434*, 167582. https://doi.org/10.1016/j.jmb.2022.167582.

42. Unsal, S., Atas, H., Albayrak, M., Turhan, K., Acar, A.C., and Doğan, T. (2022). Learning functional properties of proteins with language models. Nat. Mach. Intell. *4*, 227–245. https://doi.org/10.1038/s42256-022-00457-9.

43. Balazevic, I., Allen, C., and Hospedales, T. (2019). Multi-relational poincaré graph embeddings. Adv. Neural Inf. Process. Syst. *32*.

44. Ungar, A.A. (2001). Hyperbolic trigonometry and its application in the poincaré ball model of hyperbolic geometry. Comput. Math. Appl. *41*, 135–147. https://doi.org/10.1016/S0898-1221(01)85012-4.

45. Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., Verkuil, R., Kabeli, O., Shmueli, Y., et al. (2023). Evolutionary-scale prediction of atomic-level protein structure with a language model. Science *379*, 1123–1130. https://doi.org/10.1126/science.ade2574.

46. Han, K., Wang, Y., Chen, H., Chen, X., Guo, J., Liu, Z., Tang, Y., Xiao, A., Xu, C., Xu, Y., et al. (2023). A survey on vision transformer. IEEE Trans. Pattern Anal. Mach. Intell. *45*, 87–110. https://doi.org/10.1109/TPAMI.2022.3152247.

47. Johnson, S.R., Peshwa, M., and Sun, Z. (2024). Sensitive remote homology search by local alignment of small positional embeddings from protein language models. Elife *12*, RP91415. https://doi.org/10.7554/eLife.91415.3.

48. Chen, T., and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining, pp. 785–794. https://doi.org/10.1145/2939672.2939785.

49. Friedman, J.H. (2001). Greedy function approximation: a gradient boosting machine. Ann. Stat. *29*, 1189–1232.

50. Wei, M.-M., Yu, C.-Q., Li, L.-P., You, Z.-H., and Wang, L. (2023). BCMCMI: a fusion model for predicting circRNA-miRNA interactions combining semantic and meta-path. J. Chem. Inf. Model. *63*, 5384–5394. https://doi.org/10.1021/acs.jcim.3c00852.

51. Chen, C., Zhang, Q., Yu, B., Yu, Z., Lawrence, P.J., Ma, Q., and Zhang, Y. (2020). Improving protein-protein interactions prediction accuracy using XGBoost feature selection and stacked ensemble classifier. Comput. Biol. Med. *123*, 103899. https://doi.org/10.1016/j.compbiomed.2020.103899.

52. Zhou, H., Chen, C., Wang, M., Ma, Q., and Yu, B. (2019). Predicting golgi-resident protein types using conditional covariance minimization with XGBoost based on multiple features fusion. IEEE Access *7*, 144154–144164. https://doi.org/10.1109/ACCESS.2019.2938081.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
Phage and some Host	Database: Millard Lab	http://millardlab.org
Host	Database: Uniprot	https://www.uniprot.org
Other		
Materials	This paper	https://github.com/JIENWU/GE-PHI
Data and code	This paper	https://github.com/JIENWU/GE-PHI
Software and Algorithms	This paper	https://github.com/JIENWU/GE-PHI

METHOD DETAILS

Phage-host heterogeneous association network

It is well known that the receptor-binding proteins (RBP) on the host surface and tail proteins of phages determine whether the phage can adsorb to the target host. In this context, when constructing the microbial interaction network, we take these factors into account. Specifically, we constructed a phage-host heterogeneous association network (PHAN) consisting three components: the phage-host interaction network, the phage-phage similarity network, and the host-host similarity network. The details of each part are listed as below.

(1) Phage-Host Interaction Network

The PHI dataset used in this study includes nine bacterial hosts, comprising the ESKAPE pathogenic bacteria (*Enterococcus faecium*, *Staphylococcus aureus*, *Klebsiella pneumoniae*, *Acinetobacter baumannii*, *Pseudomonas aeruginosa*, and *Enterobacter species*), supplemented with *Clostridioides difficile*, *Salmonella enterica*, and *Escherichia coli*. The sequence information was downloaded from the Millard Lab (http://millardlab.org)[38] and the UniProt (https://www.uniprot.org)[39] database. Finally, we collected 1,232 PHI pairs consisting of 960 tail proteins of phages and 793 host RBP sequences. Let $A \in R^{N_P \times N_H}$ represented an adjacency matrix and used to describe the known PHI pairs, where N_P and N_H denotes the quantity of phages and hosts, respectively. If a phage $p_i(i \in 1, \cdots, N_P)$ has a $A_{i,j}$ demonstrated interaction with the bacterium $h_j(j, \cdots, N_H)$, the value will be set to 1; otherwise, it is set to 0.

(2) Cosine similarity network

Assuming that similar bacteria or hosts exhibit comparable functional patterns, we employed cosine similarity to construct the phage-phage and host-host similarity networks. Cosine similarity is frequently applied in various bioinformatics tasks, including gene co-expression analysis,[40] microbial community analysis,[41] and functional annotation prediction.[42] Unlike the traditional methods that focused on the length of feature vectors, cosine similarity is calculated based on their direction. Moreover, it remains valid when dealing with sparse vectors and outliers.

Cosine similarity is calculated as the inner product of two vectors divided by the product of their magnitudes, yielding values between −1 and 1. A value closer to 1 indicates higher similarity between two sequences, while a value closer to −1 suggests lower similarity. The detailed calculation of the phage-phage similarity network is as follows:

$$\cos(P_i, P_j) = \frac{\sum_{i=1}^{n}(P_i \times P_j)}{\sqrt{\sum_{i=1}^{n} P_i^2} \times \sqrt{\sum_{i=1}^{n} P_j^2}}, \quad \text{(Equation 1)}$$

where P_i and P_j represents two different phage sequences. To compute their similarity, we project these sequences into a vector space. Additionally, the cosine similarity of bacterial hosts can also be calculated in this way

In summary, based on the cosine similarity between different microbial nodes, we constructed PHAN to capture the topological information. Considering both experimental accuracy and time efficiency, the cosine similarity threshold was set to 0.9. In this way, two types of homogeneous graphs (phage-phage similarity and host-host similarity) were incorporated into the phage-host interaction network. Subsequently, known phage-host associations were used to connect the two homogeneous graphs, forming the final heterogeneous graph. In PHAN, the feature representations of microbial nodes are derived from their functional similarity. Additionally, the attribute representations were derived from the adjacency matrix.

Hyperbolic graph embeddings for PHAN

Graph embedding techniques are capable of uncovering hidden information and non-linear patterns among edges and entities. However, conventional graph embedding methods are inadequate for our research, as they primarily focus on connections between biological nodes while neglecting the attributes of edges and entities. After constructing PHAN, a knowledge graph embedding (KGE) algorithm, Multi-relational Poincaré Graph Embeddings (MuRP),[43] was employed to capture low-rank representations for all entities and relations.

Microbial knowledge graphs are frequently incomplete, making link prediction essential for inferring true relationships. Therefore, we focus on embedding multi-relational microbial nodes within hyperbolic space. Let a $Graph = (e_s, r, e_o) \in E \times R \times E$, where E denotes the set of microbial nodes, and R represents the set of associated edges. Each entity $e_s \in E$ and relation $r \in R$ is embedded into the Poincaré ball B_c^d, and then the space can be defined by:

$$B_c^d = \{ y \in R^d : c \, \| \, y \| \; < 1 \}. \qquad \text{(Equation 2)}$$

The hyperbolic distance between two e_s and e_o in the space (shortest path between two nodes, seen in Figure 9A) can be calculated as follows:

$$d_{B_c}(e_s, e_o) = \frac{2}{\sqrt{c}} \tanh^{-1}\left(\sqrt{c} \, \| - e_s \oplus_c e_o \|\right), \qquad \text{(Equation 3)}$$

where $\| \cdot \|$ describes the Euclidean norm. $\oplus_c$ represents the *Möbius addition*,[44] which is a non-Euclidean vector operation critical for preserving the geometric properties of the Poincaré ball. The detailed *Möbius addition* can be defined as:

$$e_s \oplus_c r = \frac{(1+2c\langle e_s, r\rangle + c\|r\|^2)e_s + (1 - c\|e_s\|^2)r}{1+2c\langle e_s, r\rangle + c^2\|e_s\|^2\|r\|^2}, \qquad \text{(Equation 4)}$$

where $\langle \cdot, \cdot \rangle$ represents the Eucliden inner product. This distance metric is well-suited to capture the hierarchical nature of relationships between microbial nodes. In the model, the embedding of a relation r adjusts the distance between nodes, allowing the model to represent multi-relational data effectively within the Poincaré ball. The score function of MuRP is

$$\phi_{MuRP} = - d_{B_c}\left(\exp_o^c\left(R\log_o^c(e_s)\right), e_o \oplus_c r\right)^2 + b_s + b_o. \qquad \text{(Equation 5)}$$

Here, e_s and e_o represents the embeddings of the source and target nodes, r denotes the embeddings of their relationships, b_s and b_o are the bias term. The optimization objective of the MuRP to optimize the Bernoulli negative log-likhood loss can be calculated as:

$$L = - \sum_{(e_s, r, e_o) \in O} \log \sigma(\delta - d_{B_c}(e_s \oplus_c r, e_o)) + \mu \|r\|^2. \qquad \text{(Equation 6)}$$

In Equation 6, σ is the sigmoid function, δ represents a margin hyper-parameter, and μ is a regularization term to avoid overfitting. After the above steps, we can extract the low-dimensional representations of PHAN and capture the complex associations.

Evolutionary scale modeling of sequence information

In bioinformatics, natural language processing (NLP) techniques can learn the underlying grammar of biological sequences by training on universal proteome databases. After utilizing the MuRP model to capture the low-dimensional topological information from PHAN, we introduced the pre-trained ESM-2 (t33_650M_UR50D)[45] protein language model to learn the biological properties from phage tail proteins and host RBP sequences. ESM-2 is a state-of-the-art protein language model that leverages large-scale evolutionary data to generate high-dimensional representations of protein sequences and captures both functional and structural characteristics.

The core of ESM-2 is the transformer architecture,[46] which consists of multiple levels of the self-attention mechanism and a feed-forward neural network (Figure 1C). The self-attention mechanism allows the model to account for positional relationships as it processes sequences, while the feedforward neural network further integrates this information. More specifically, ESM-2 utilized the multiple-head self-attention mechanism to capture long-rang dependencies within trail protein and RBP sequences. This mechanism captures global sequence information by calculating the correlation of each position in the sequence with all other positions. In addition, ESM-2 also introduced the encoding of evolutionary information in the model, allowing the model to learn both conservation and variability of protein sequences to improve the prediction accuracy. To maintain the order information of elements in sequences, ESM-2 incorporates positional encoding, which helps understand the relative relationship between different position.[47] The final output is a prediction of the 3D structure of target protein, expressed as atomic coordinates. These coordinates describe the spatial location of atoms in the protein molecule and can be applied for further biophysical analyses.

Machine learning-based classifier

XGBoost (eXtreme Gradient Boosting)[48] is an ensemble learning algorithm based on Gradient Boosting Decision Trees (GBDT)[49] and can be used for both classification and regression tasks. This technique has been widely applied in various bioinformatics fields, such as predicting circRNA-miRNA interactions,[50] protein-protein interactions,[51] and Golgi-resident protein types.[52] Unlike traditional

GBDT that rely solely on first-order derivatives for loss function optimization, XGBoost incorporates second-order derivatives to enhance the customization of the loss function, thereby improving both accuracy and speed. It also adds Regularization terms to the loss function to control model complexity and prevent overfitting.

For a given PHI dataset $D = \{[H_n, y_n]\}$, where $|D| = N$ and $H_n \in R^m$ represents the concatenated feature vector for each entity $n \in N$, which were derived from MuRP and ESM-2, and $y_n \in R$ denotes the target label. The prediction process can be described as follows:

$$\hat{y}_n = \sum_{k=1}^{K} f_k(H_n), \qquad \text{(Equation 7)}$$

where k represents the total number of trees, and $f_k(H_n)$ denotes the prediction score of the sample H_n on the k-th tree. XGBoost utilizes regression trees as its base classifier functions, and thus the prediction scores can be calculated as:

$$Obj(\theta) = \sum_{n=1}^{N} L(y_n, \widehat{y}_n) + \sum_{k=1}^{K} \Omega(f_k), \qquad \text{(Equation 8)}$$

where $L(y_n, \widehat{y}_n)$ denotes the loss function measuring the error between the true label y_n and the predicted label $\widehat{y}_n$. $\Omega(f_k)$ is the regularization term that controls the complexity of k-th tree. The prediction results of the combined t tree models on the sample H_n is updated as follows:

$$\widehat{y}_n^{(t)} = \widehat{y}_n^{(t-1)} + f_t(H_n). \qquad \text{(Equation 9)}$$

The term $\Omega(f_k)$ describes the complexity of the k-th tree and can be expressed as:

$$\Omega(f_k) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2. \qquad \text{(Equation 10)}$$

Here, γ and λ is the regularization parameters, and w_j represents the weight of the leaf nodes in the trees. The model can be expressed as

$$f_t(H_n) = w_{q(H_n)}, w \in R^T, \qquad \text{(Equation 11)}$$

where $q(H_n)$ indicates the leaf node of H_n, and T is the number of leaf nodes in the tree. The first derivative g_i and the second derivative h_i are used to approximate the objective function through Taylor expansion:

$$g_i = \sum_{i \in L_j} g_i = \sum_{i \in L_j} \partial_{\hat{y}_i^{t-1}} l\left(y_i, \widehat{y}_i^{(t-1)}\right), \qquad \text{(Equation 12)}$$

$$h_j = \sum_{i \in L_j} h_i = \sum_{i \in L_j} \partial^2_{\hat{y}_i^{(t-1)}} l\left(y_i, \widehat{y}_i^{(t-1)}\right). \qquad \text{(Equation 13)}$$

The objective function can be transformed into the leaf node form of the t-th tree by combining the above formulas and is expressed as

$$Obj^{(t)}(\theta) \approx \sum_{j=1}^{T} \left[g_j w_j + \frac{1}{2}(h_j + \lambda) w_j^2 \right] + \gamma T. \qquad \text{(Equation 14)}$$

The optimal weight w_j^* for each leaf node are obtained by

$$w_j^* = -\frac{g_i}{h_j + \lambda}. \qquad \text{(Equation 15)}$$

Evaluation metrics

In the experiments, we assessed the performance of the GE-PHI model using a 5-fold cross-validation (CV). Specifically, the PHI dataset was partitioned into five equal folds, with each fold being alternately used as the test set, while the remaining folds served as the training set. This process was repeated until each part had been used as validation set once. To comprehensively evaluate the model's performance, we employed several metrics, including Accuracy (Acc), F1-score (F1), Precision (Pre), Recall (Rec), and Matthew's correlation coefficient (MCC). The definitions of these metrics are as follows:

$$Accuracy = \frac{TN + TP}{TN + FP + FN + TP}, \qquad \text{(Equation 16)}$$

$$F1 - scoure = \frac{precision \times recall}{recall + precision} \times 2, \quad \text{(Equation 17)}$$

$$\text{Precison} = \frac{TP}{FN + TP}, \quad \text{(Equation 18)}$$

$$\text{Recall} = \frac{TP}{FP + TP}, \quad \text{(Equation 19)}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP) \times (TP + FN) \times (TN + FN) \times (TN + FP)}}. \quad \text{(Equation 20)}$$

where true positives (TP) and true negatives (TN) refer to correctly predicted positive and negative PHI samples, respectively. Likewise, false positives (FP) and false negatives (FN) represent misclassified samples. Additionally, the Area Under the Receiver Operating Characteristic (ROC) curve (AUC) and the Area Under the Precision-Recall (PR) curve (AUPR) were used to assess the GE-PHI model's predictive performance. AUC measures the model's ability to distinguish between positive and negative samples across different threshold settings, whereas AUPR emphasizes the trade-off between precision and recall, especially in imbalanced datasets. Higher values of AUC and AUPR indicate better model performance.

QUANTIFICATION AND STATISTICAL ANALYSIS

The evaluation metrics were calculated using scikit-learn version 0.19.0.