Supplementary Material for ‘A novel framework for predicting phage-host interactions via host specificity-aware graph autoencoder’

1. Summary of dataset across different host taxa

To provide a more intuitive representation of the interactions between phages and hosts across different taxonomic groups, we counted the number and frequency of phages infecting varying numbers of hosts within different taxonomic host groups, as detailed in Table S1.

Table S1. Statistics on the number and frequency of phages infecting different host taxa. The number in parentheses has been multiplied by 100, which represents the proportion.

The number of hosts in different taxa	Host taxa					
	Species	Genus	Family	Order	Class	Phylum
1	1950 (99.29)	1962 (99.90)	1963 (99.95)	1963 (99.95)	1964 (100.00)	1964 (100.00)
2	13 (0.66)	1 (0.05)	1 (0. 05)	1 (0. 05)	0 (0. 00)	0 (0. 00)
3	1 (0. 05)	1 (0. 05)	0 (0. 00)	0 (0. 00)	0 (0. 00)	0 (0. 00)

From Table S1, it is evident that in our dataset, over 99% of bacteriophages exclusively infect a single bacterial species. Moreover, at the genus level, only 0.1% of bacteriophages exhibit inter-genus infectivity. This observation distinctly demonstrates the strong host specificity of bacteriophages.

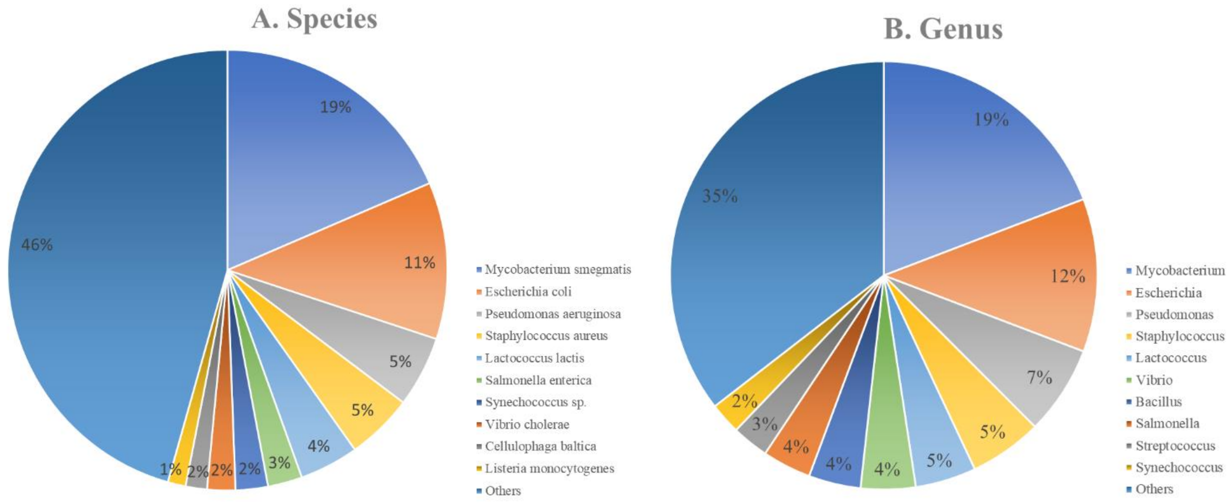


Figure S1. The distribution of host species (A) and genus (B) for our dataset.

2. The distribution of DNA/protein sequence lengths of phages

We extracted sequence features based on the whole genome sequence and protein sequence of phages and then calculated the similarity between them. The whole genome sequence length of each phage is different, and there are often multiple protein sequences for each phage. Therefore, we select several features that are not related to sequence length. Figure S2 shows the length distribution of DNA and protein sequences of phages.

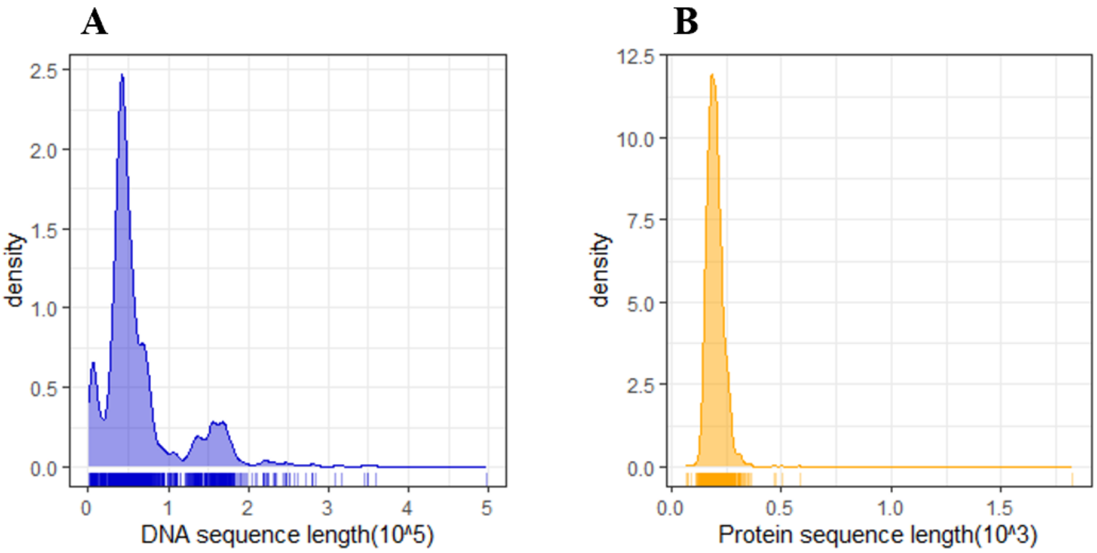


Figure S2. The distribution of DNA and protein sequence lengths of phages. A is the distribution of DNA sequences lengths, B is the distribution of protein sequences lengths. (x-axis represent the sequence length, y-axis represent the density and the scale values are multiplied by 1000.)

As shown in the figureS1, we can find that the vast majority of phage DNA sequence lengths are less than 105, a few are between 105 and 2×105, and almost rarely above 2×105. There is more than one protein sequence of phage, and we take the average of all the sequence lengths, which are mostly within 0.5×103.

3. Feature extraction of DNA sequences and protein sequences of phages

Table S2. Features of DNA and protein sequences for constructing phage-phage similarities.

Molecule	Feature	Interpretation
DNA	k-mer	the occurrence of a sequence of k adjacent nucleic acids
	NAC	the frequency of each type of nucleic acid composing a nucleotide sequence
Protein	AAC	the frequency of each type of amino acid composing a protein sequence
	AC	the abundance of specific chemical elements constituting a protein sequence
	MW	molecular weight of the protein sequence

For a given DNA sequence, there are often multiple feature extraction methods that can be used, such as k-mer, RCKmer, Nucleotide composition (NAC), DNC, and others. In our study, we chose the two most commonly used feature extraction methods, k-mer and NAC, to represent the features of the DNA sequence.

The feature k-mer frequency refers to the occurrence frequencies of k adjacent nucleotides in a DNA sequence, which is simply denoted as k-mer. Originally, k-mers refers to the sequence fragments containing k adjacent nucleotides. If a nucleic acid sequence is L in length, then the number of k-mers is L-k+1. The k-mer frequency is calculated by selecting a value of k, decomposing the DNA sequence into all possible k-mers by sliding one nucleotide at a time and calculating the frequency of each type of k-mer, finally constructing a vector containing all possible k-mers frequencies in the sequence. The k-mer (k = 4) is formulated as follows:

$$\text{k-mer} = N(m)/N_k, \quad m \in \{AAAA, AAAC, AAAG, AAAT, ..., TTTT\},$$

where *N(m)* is the number of each type of k-mer, N_k is the number of all *k*-mers in the nucleotide sequence.

Nucleotide composition (NAC) represents the occurrence frequency of each nucleotide in a nucleotide sequence, where these frequencies form a feature vector. Considering the DNA sequence as a sequence composed of four nucleotides (*A, T, C, G*), the frequencies of the four nucleotides can be calculated as:

$$NAC = N(t)/N,\ \ t \in \{A, C, G, T\},$$

where *N(t)* is the number of each nucleotide type, *N* is the length of the DNA sequence.

For amino acid sequences, we selected three popular feature representations, namely, amino acid composition (AAC), chemical element abundance composition (AC), and molecular weight (MW).

Amino acid composition (AAC) is a representation of the occurrence frequency of each amino acid in a protein sequence. All amino acids (ACDEFGHIKLMNPQRSTVWY*) contain 20 natural amino acids and other unknown amino acids. The AAC can be calculated as:

$$AAC = N(t)/N,\ \ t \in \{A, C, D, ..., Y, *\},$$

where *N(t)* is the number of amino acid *t*. *N* is the total number of amino acids in a protein sequence.

Chemical element abundance composition (AC) represents the occurrence frequencies of five chemical elements (C, H, O, N, S) in a protein sequence, which can be formulated as:

$$AC = N(t)/N,\ \ t \in \{C, H, O, N, S\}.$$

N(t) is the number of chemical element *t*, and *N* is the length of the protein sequence.

The molecular weight (MW) of the protein represents the molecular weight of a protein sequence, which can be formulated as:

$$MW = \sum w(t) - (\mathrm{L}-1) \times 18.01,\ \ t \in \{A, C, ..., Y, *\}.$$

w(t) is the molecular weight of amino acid *t*, and *L* is the length of the protein sequence.

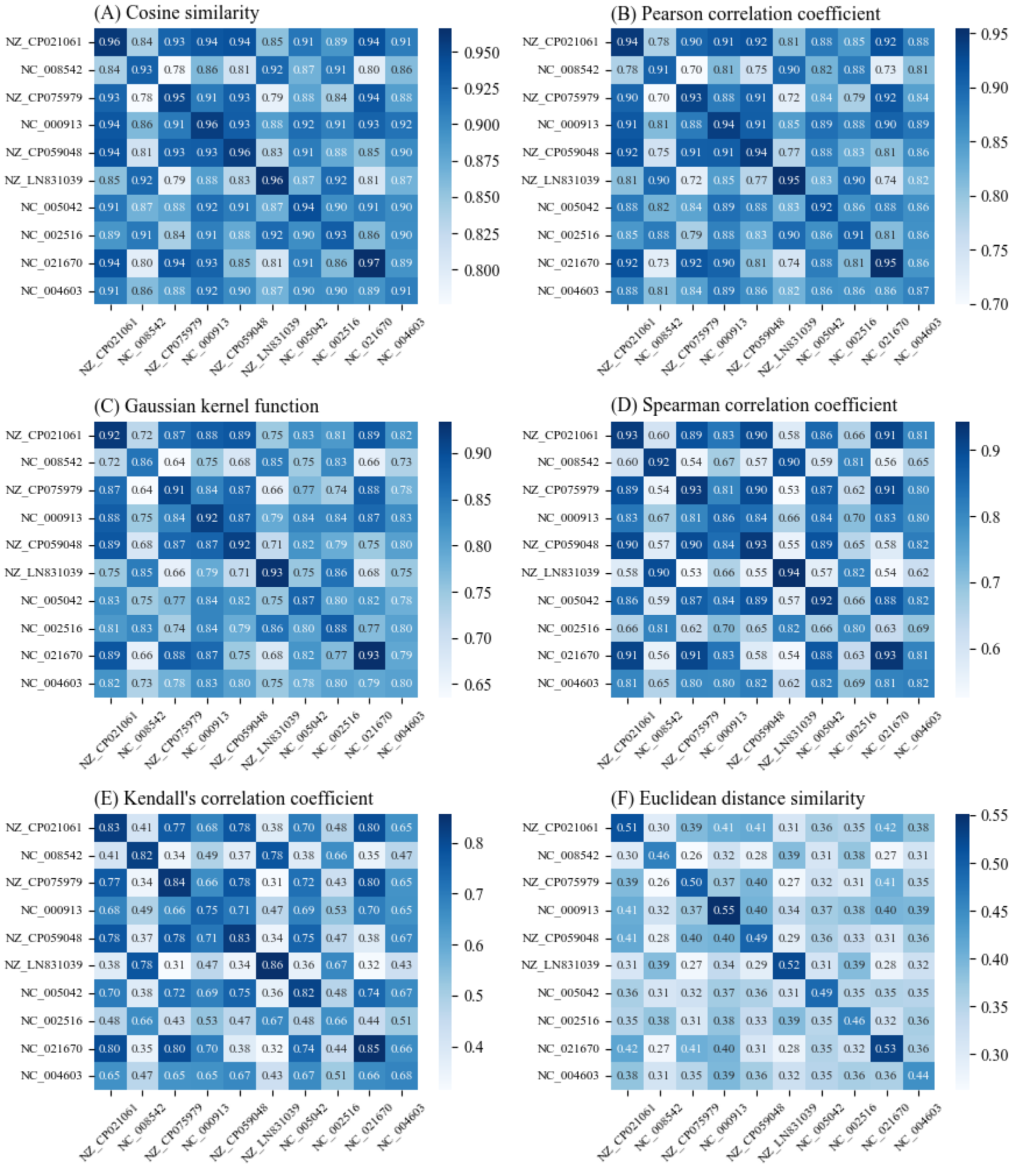
4. The effect of various similarity measures

To pick an appropriate similarity measure, we selected 10 bacteria at the species level and 10 phages associated with each from our dataset, for a total of 100 phage-host interactions. The bacteria names and their respective NCBI accession numbers are shown in Table S3. We use various similarity measures (Cosine similarity, Pearson correlation coefficient, Gaussian kernel function, Spearman correlation coefficient, Kendall's correlation coefficient, and Euclidean distance) to calculate the similarity between phages, expecting that phages with the same host tend to be more similar, which is also consistent with the premise assumption in our model.

As can be seen from Figure S3, the results obtained by the Gaussian kernel function showed little difference between phages associated with different bacteria. On the contrary, the results obtained by Euclidian distance similarity are excessively different, and the similarity between phages corresponding to the same host is also small, which is not conducive to the subsequent construction of local connected graphs to explore important local relationships. Other results fall somewhere in between. The results calculated by cosine similarity and Pearson correlation coefficient are more consistent with our expectation that there is greater similarity between phages associated with the same host. Therefore, we finally chose the more commonly used cosine similarity as the similarity calculation method between phages.

Table S3. The bacteria names and their respective NCBI accession numbers

Bacteria name	NCBI accession number
Bacillus thuringiensis	NZ_CP021061
Burkholderia cenocepacia	NC_008542
Clostridium perfringens	NZ_CP075979
Escherichia coli	NC_000913
Lactococcus lactis	NZ_CP059048
Mycobacterium smegmatis	NZ_LN831039
Prochlorococcus marinus	NC_005042
Pseudomonas aeruginosa	NC_002516
Staphylococcus aureus	NC_021670
Vibrio parahaemolyticus	NC_004603



(A) Cosine similarity

	NZ_CP021061	NC_008542	NZ_CP075979	NC_000913	NZ_CP059048	NZ_LN831039	NC_005042	NC_002516	NC_021670	NC_004603
NZ_CP021061	0.96	0.84	0.93	0.94	0.94	0.85	0.91	0.89	0.94	0.91
NC_008542	0.84	0.93	0.78	0.86	0.81	0.92	0.87	0.91	0.80	0.86
NZ_CP075979	0.93	0.78	0.95	0.91	0.93	0.79	0.88	0.84	0.94	0.88
NC_000913	0.94	0.86	0.91	0.96	0.93	0.88	0.92	0.91	0.93	0.92
NZ_CP059048	0.94	0.81	0.93	0.93	0.96	0.83	0.91	0.88	0.85	0.90
NZ_LN831039	0.85	0.92	0.79	0.88	0.83	0.96	0.87	0.92	0.81	0.87
NC_005042	0.91	0.87	0.88	0.92	0.91	0.87	0.94	0.90	0.91	0.90
NC_002516	0.89	0.91	0.84	0.91	0.88	0.92	0.90	0.93	0.86	0.90
NC_021670	0.94	0.80	0.94	0.93	0.85	0.81	0.91	0.86	0.97	0.89
NC_004603	0.91	0.86	0.88	0.92	0.90	0.87	0.90	0.90	0.89	0.91

(B) Pearson correlation coefficient

	NZ_CP021061	NC_008542	NZ_CP075979	NC_000913	NZ_CP059048	NZ_LN831039	NC_005042	NC_002516	NC_021670	NC_004603
NZ_CP021061	0.94	0.78	0.90	0.91	0.92	0.81	0.88	0.85	0.92	0.88
NC_008542	0.78	0.91	0.70	0.81	0.75	0.90	0.82	0.88	0.73	0.81
NZ_CP075979	0.90	0.70	0.93	0.88	0.91	0.72	0.84	0.79	0.92	0.84
NC_000913	0.91	0.81	0.88	0.94	0.91	0.85	0.89	0.88	0.90	0.89
NZ_CP059048	0.92	0.75	0.91	0.91	0.94	0.77	0.88	0.83	0.81	0.86
NZ_LN831039	0.81	0.90	0.72	0.85	0.77	0.95	0.83	0.90	0.74	0.82
NC_005042	0.88	0.82	0.84	0.89	0.88	0.83	0.92	0.86	0.88	0.86
NC_002516	0.85	0.88	0.79	0.88	0.83	0.90	0.86	0.91	0.81	0.86
NC_021670	0.92	0.73	0.92	0.90	0.81	0.74	0.88	0.81	0.95	0.86
NC_004603	0.88	0.81	0.84	0.89	0.86	0.82	0.86	0.86	0.86	0.87

(C) Gaussian kernel function

	NZ_CP021061	NC_008542	NZ_CP075979	NC_000913	NZ_CP059048	NZ_LN831039	NC_005042	NC_002516	NC_021670	NC_004603
NZ_CP021061	0.92	0.72	0.87	0.88	0.89	0.75	0.83	0.81	0.89	0.82
NC_008542	0.72	0.86	0.64	0.75	0.68	0.85	0.75	0.83	0.66	0.73
NZ_CP075979	0.87	0.64	0.91	0.84	0.87	0.66	0.77	0.74	0.88	0.78
NC_000913	0.88	0.75	0.84	0.92	0.87	0.79	0.84	0.84	0.87	0.83
NZ_CP059048	0.89	0.68	0.87	0.87	0.92	0.71	0.82	0.79	0.75	0.80
NZ_LN831039	0.75	0.85	0.66	0.79	0.71	0.93	0.75	0.86	0.68	0.75
NC_005042	0.83	0.75	0.77	0.84	0.82	0.75	0.87	0.80	0.82	0.78
NC_002516	0.81	0.83	0.74	0.84	0.79	0.86	0.80	0.88	0.77	0.80
NC_021670	0.89	0.66	0.88	0.87	0.75	0.68	0.82	0.77	0.93	0.79
NC_004603	0.82	0.73	0.78	0.83	0.80	0.75	0.78	0.80	0.79	0.80

(D) Spearman correlation coefficient

	NZ_CP021061	NC_008542	NZ_CP075979	NC_000913	NZ_CP059048	NZ_LN831039	NC_005042	NC_002516	NC_021670	NC_004603
NZ_CP021061	0.93	0.60	0.89	0.83	0.90	0.58	0.86	0.66	0.91	0.81
NC_008542	0.60	0.92	0.54	0.67	0.57	0.90	0.59	0.81	0.56	0.65
NZ_CP075979	0.89	0.54	0.93	0.81	0.90	0.53	0.87	0.62	0.91	0.80
NC_000913	0.83	0.67	0.81	0.86	0.84	0.66	0.84	0.70	0.83	0.80
NZ_CP059048	0.90	0.57	0.90	0.84	0.93	0.55	0.89	0.65	0.58	0.82
NZ_LN831039	0.58	0.90	0.53	0.66	0.55	0.94	0.57	0.82	0.54	0.62
NC_005042	0.86	0.59	0.87	0.84	0.89	0.57	0.92	0.66	0.88	0.82
NC_002516	0.66	0.81	0.62	0.70	0.65	0.82	0.66	0.80	0.63	0.69
NC_021670	0.91	0.56	0.91	0.83	0.58	0.54	0.88	0.63	0.93	0.81
NC_004603	0.81	0.65	0.80	0.80	0.82	0.62	0.82	0.69	0.81	0.82

(E) Kendall's correlation coefficient

	NZ_CP021061	NC_008542	NZ_CP075979	NC_000913	NZ_CP059048	NZ_LN831039	NC_005042	NC_002516	NC_021670	NC_004603
NZ_CP021061	0.83	0.41	0.77	0.68	0.78	0.38	0.70	0.48	0.80	0.65
NC_008542	0.41	0.82	0.34	0.49	0.37	0.78	0.38	0.66	0.35	0.47
NZ_CP075979	0.77	0.34	0.84	0.66	0.78	0.31	0.72	0.43	0.80	0.65
NC_000913	0.68	0.49	0.66	0.75	0.71	0.47	0.69	0.53	0.70	0.65
NZ_CP059048	0.78	0.37	0.78	0.71	0.83	0.34	0.75	0.47	0.38	0.67
NZ_LN831039	0.38	0.78	0.31	0.47	0.34	0.86	0.36	0.67	0.32	0.43
NC_005042	0.70	0.38	0.72	0.69	0.75	0.36	0.82	0.48	0.74	0.67
NC_002516	0.48	0.66	0.43	0.53	0.47	0.67	0.48	0.66	0.44	0.51
NC_021670	0.80	0.35	0.80	0.70	0.38	0.32	0.74	0.44	0.85	0.66
NC_004603	0.65	0.47	0.65	0.65	0.67	0.43	0.67	0.51	0.66	0.68

(F) Euclidean distance similarity

	NZ_CP021061	NC_008542	NZ_CP075979	NC_000913	NZ_CP059048	NZ_LN831039	NC_005042	NC_002516	NC_021670	NC_004603
NZ_CP021061	0.51	0.30	0.39	0.41	0.41	0.31	0.36	0.35	0.42	0.38
NC_008542	0.30	0.46	0.26	0.32	0.28	0.39	0.31	0.38	0.27	0.31
NZ_CP075979	0.39	0.26	0.50	0.37	0.40	0.27	0.32	0.31	0.41	0.35
NC_000913	0.41	0.32	0.37	0.55	0.40	0.34	0.37	0.38	0.40	0.39
NZ_CP059048	0.41	0.28	0.40	0.40	0.49	0.29	0.36	0.33	0.31	0.36
NZ_LN831039	0.31	0.39	0.27	0.34	0.29	0.52	0.31	0.39	0.28	0.32
NC_005042	0.36	0.31	0.32	0.37	0.36	0.31	0.49	0.35	0.35	0.35
NC_002516	0.35	0.38	0.31	0.38	0.33	0.39	0.35	0.46	0.32	0.36
NC_021670	0.42	0.27	0.41	0.40	0.31	0.28	0.35	0.32	0.53	0.36
NC_004603	0.38	0.31	0.35	0.39	0.36	0.32	0.35	0.36	0.36	0.44

Figure S3. Similarity between 100 phages associated with ten bacteria at the species level. Subgraphs A to F are the results of Cosine similarity, Pearson correlation coefficient, Gaussian kernel function, Spearman correlation coefficient, Kendall's correlation coefficient, and Euclidean distance similarity, respectively.

5. The effect of the choice of k in k-mers on the prediction performance

We used the k-mer frequency to extract the DNA sequence features of phage. To evaluate the effect of different k on the model performance, we select several common values of k and use 5-fold cross-validation to compared the model performances. Since our aim here is to choose the best k-value, we use only DNA sequence features including k-mer frequencies for the similarity between phages to assess the effect of different k-values on the model performance.

As can be seen from the results in Figure S4, there is no obvious difference in the evaluation metrics AUC and accuracy of the model based on different values of k at both species level and genus level, which also indicates that the value of k does not greatly affect the similarity between phages. In other words, the value of k does not have a significant impact on the prediction performance of the model. Therefore, considering the impact of the dimensionality of the features on the computational time complexity, we chose k = 4 to calculate the k-mer frequency.

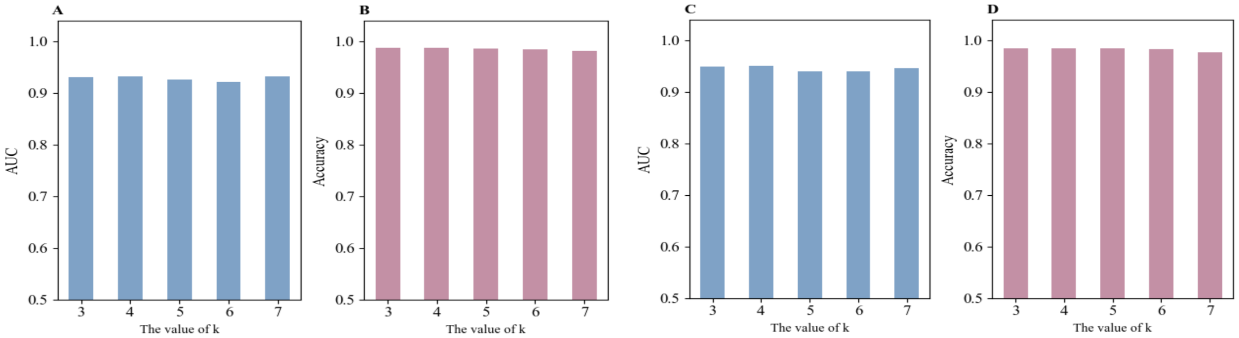


Figure S4. The prediction performance of models based on different k in k -mers under 5-fold cross-validation (x-axis represents different values of k). A and B are AUC and accuracy based on different k values at the species level, respectively. C and D are AUC and accuracy based on different k values at the genus level, respectively.

6. The derivation of loss function in optimization process

The original matrix is denoted as *X*, where m and n represent the number of rows and columns of the matrix *X*, respectively. To encourage the model to produce row vectors with fewer non-zero elements when reconstructing the association matrix, we consider sparse regularization of each row x_i of the matrix by introducing sparse penalty terms such that as many zeros as possible occur in each row. The objective function is defined as follows:

$$\min_{X} L(X)+\lambda\sum_{i=1}^{m}\left(\|x_i\|_1-\varepsilon\right),$$

where λ is the hyperparameter that controls the intensity of sparse penalty, ε is the sparsity parameter and $L(X)$ is the loss function of the original matrix *X*, such as the cross-entropy loss function in this paper.

We then show that by introducing this penalty, the number of non-zero elements per row in the optimization problem will be reduced, resulting in more zeros per row in the original matrix.

Assuming that the optimal solution of the original optimization problem is X^*, and the optimal solution after adding sparse regularization is X^{**}, then we have:

$$L\left(X^{**}\right)+\lambda\sum_{i=1}^{m}\left(\left\|x_i^{**}\right\|_1-\varepsilon\right)\leq L\left(X^{*}\right)+\lambda\sum_{i=1}^{m}\left(\left\|x_i^{*}\right\|_1-\varepsilon\right).$$

Let $\Delta L = L\left(X^{**}\right)-L\left(X^{*}\right)$, the above formula can be further simplified as:

$$\Delta L \leq \lambda\sum_{i=1}^{m}\left(\left\|x_i^{*}\right\|_1-\left\|x_i^{**}\right\|_1\right).$$

Since all the elements of the matrix are non-negative, so $\left\|x_i^{*}\right\|_1-\left\|x_i^{**}\right\|_1\geq 0$. Hence,

$$\Delta L \le \lambda \sum_{i=1}^{m} \left(\left\| x_i^* \right\|_1 - \left\| x_i^{**} \right\|_1 \right) \le \lambda m \max_i \left(\left\| x_i^* \right\|_1 - \left\| x_i^{**} \right\|_1 \right).$$

According to the triangle inequality, we have

$$\left\| x_i^{**} \right\|_1 \ge \left\| x_i^* \right\|_1 - \left\| x_i^{**} - x_i^* \right\|_1.$$

Then we can obtain

$$\begin{aligned} \Delta L &\le \lambda m \max_i \left(\left\| x_i^* \right\|_1 - \left\| x_i^{**} \right\|_1 \right) \\ &\le \lambda m \max_i \left(\left\| x_i^* - x_i^{**} \right\|_1 \right) \\ &\le \lambda m \max_i \sum_j \left| x_{ij}^* - \operatorname{sign}\left(x_{ij}^{**} \right) \right| \end{aligned},$$

where

$$\operatorname{sign}(x) = \begin{cases} 1, & x > 0 \\ 0, & x = 0 \\ -1, & x < 0 \end{cases}.$$

For $\max_i \sum_j \left| x_{ij}^* - \operatorname{sign}\left(x_{ij}^{**} \right) \right|$, it can be shown that it obtains the minimum value when $\operatorname{sign}\left(x_{ij}^{**} \right) = 0$. Since $\left| x_{ij}^* - \operatorname{sign}\left(x_{ij}^{**} \right) \right| \ge 1$ when $\operatorname{sign}\left(x_{ij}^{**} \right) = 1$ or $\operatorname{sign}\left(x_{ij}^{**} \right) = -1$, then we have

$$\Delta L \le \lambda m \max_i \sum_j \left| x_{ij}^* - \operatorname{sign}\left(x_{ij}^{**} \right) \right| \le \lambda m \max_i \sum_j \left| x_{ij}^* \right|.$$

In the above equation, $\max_i \sum_j \left| x_{ij}^* \right|$ denotes the number of non-zero elements in each row of the original matrix, because $\Delta L = L\left(X^{**} \right) - L\left(X^* \right) < 0$, so $\max_i \sum_j \left| x_{ij}^* \right| > \max_i \sum_j \left| x_{ij}^{**} \right|$, that is, after adding the sparse regular term, each row of the matrix will generate more zeros. The proof is complete.

7. Hyperparameter sensitivity analysis

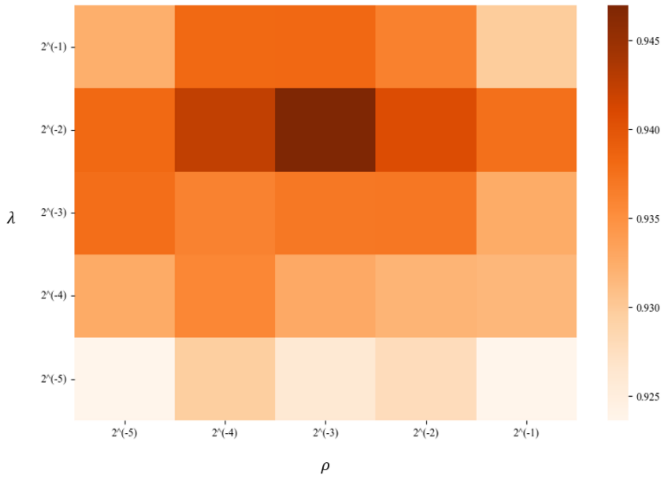


Figure S5. AUC of models with different $\boldsymbol{\lambda}$ and $\boldsymbol{\rho}$.

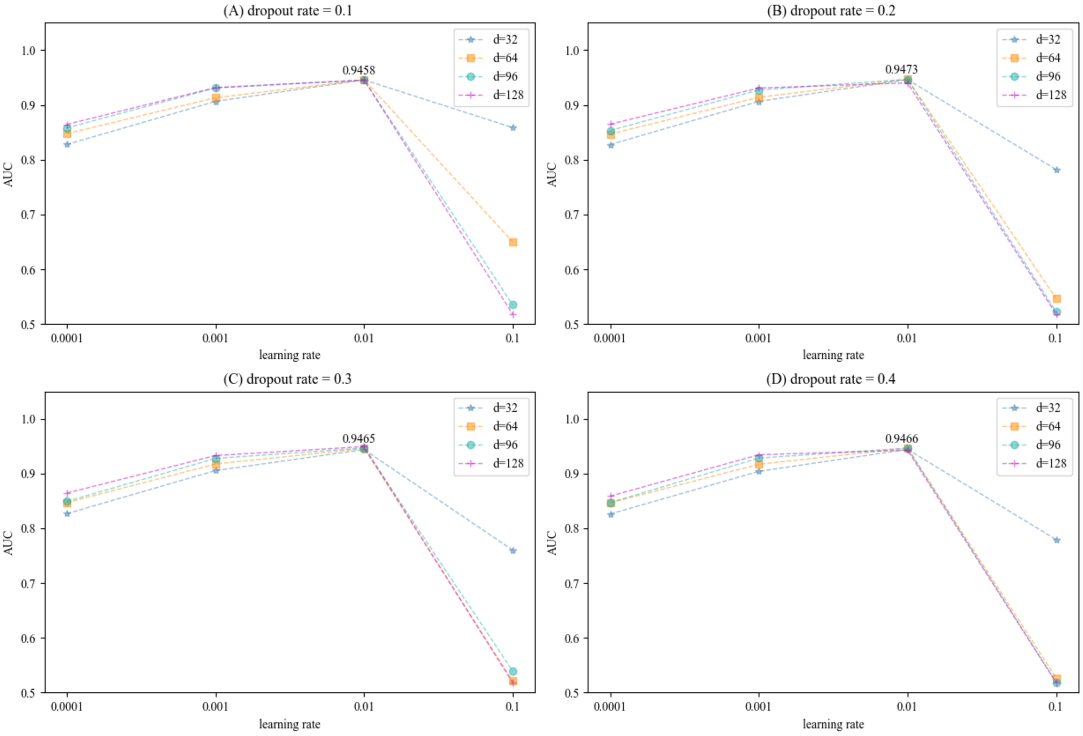


Figure S6. Hyperparameter sensitivity analysis of PHISGAE

8. The significance test for the trends on *K*

In order to prove that the performance variation trend is affected by *K* and is statistically significant in the two ranges of *K*, we first conducted one-way ANOVA respectively, and the results are shown in Table S4. The P-values under both scenarios are less than 0.05, indicating that changes in the results under different ranges of *K* are statistically significant. Further, we conducted multiple comparative analyses respectively to test the statistical significance of performance trends with different *K* values, and results are shown in Table S5-S6.

Table S4. The ANOVA results.

N	F	P-value
10-100	89.92	4.5e-41
3-10	16.58	8.02e-13

We used the Bonferroni correction method commonly used in multiple comparison analysis to correct the P-values of multiple comparisons. Combined with Figure 3 in the text, it can be seen from Table S5 that the performance decline trend is basically statistically significant as *K* goes from 10 to 100. Although the trend in performance for *K* between 3 and 10 is not exactly statistically significant, it can be found that only when *K* is set to 6, the results are significantly variable compared to both ends of its range, and given Figure 5 in the main text, we can also determine that 6 is the optimal choice for *K*.

Table S5. The P-value of pairwise sample mean test in multiple comparison analysis. (*K* =10, 20, …, 100)

K	10	20	30	40	50	60	70	80	90
20	0.00106	-	-	-	-	-	-	-	-
30	4.3e-09	0.03549	-	-	-	-	-	-	-
40	1.2e-15	2.8e-07	0.02287	-	-	-	-	-	-

50	< 2e-16	8.0e-11	4.1e-05	0.50978	-	-	-	-	-
60	< 2e-16	< 2e-16	9.4e-13	1.8e-06	0.00200	-	-	-	-
70	< 2e-16	< 2e-16	2.7e-14	8.1e-08	0.00014	1.00000	-	-	-
80	< 2e-16	< 2e-16	< 2e-16	1.8e-10	5.9e-07	0.34741	1.00000	-	-
90	< 2e-16	< 2e-16	< 2e-16	2.6e-10	8.3e-07	0.37917	1.00000	1.00000	-
100	< 2e-16	< 2e-16	< 2e-16	1.4e-12	6.2e-09	0.02788	0.19156	1.00000	1.00000

Table S6. The P-value of pairwise sample mean test in multiple comparison analysis. (K =3, 4, …,10)

K	3	4	5	6	7	8	9
4	4.7e-07	-	-	-	-	-	-
5	2.6e-10	1.000	-	-	-	-	-
6	1.2e-12	0.062	1.000	-	-	-	-
7	6.9e-11	0.576	1.000	1.000	-	-	-
8	1.0e-08	1.000	1.000	0.576	1.000	-	-
9	9.9e-08	1.000	1.000	0.165	1.000	1.000	-
10	7.0e-06	1.000	0.278	0.008	0.131	1.000	1.000

9. Other comparisons of prediction performance.

Table S7. The performance of PHISGAE and other methods based on 10 repetitions of 5-fold cross-validation.

Methods	AUPR	AUC	ACC	MCC
LAGCN	0.3254±0.0024	0.8033±0.0019	0.9208±0.0014	0.1353±0.0011
PHIAF	0.4442±0.0126	0.8638±0.0034	0.7965±0.0056	0.1481±0.0452
NIMGSA	0.4122±0.0111	0.8738±0.0020	0.9974±0.0010	0.4376±0.0295
PredPHI	0.4361±0.0207	0.9035±0.0053	0.8166±0.0017	0.3457±0.1112
ILMF-VH	0.4042±0.0081	0.9088±0.0019	0.9982±0.0011	0.4296±0.0186
CHERRY	0.3937±0.0026	0.9178±0.0016	0.9698±0.0017	0.2255±0.0063
PHISGAE	0.4547±0.0038	0.9467±0.0015	0.9880±0.0006	0.4218±0.0085

Table S8. The performance of PHISGAE and other methods on independent dataset.

Methods	AUPR	AUC	ACC	MCC
LAGCN	0.3220	0.8005	0.9389	0.1367
PHIAF	0.4265	0.8621	0.8263	0.1411
NIMGSA	0.4106	0.8750	0.9931	0.4209
PredPHI	0.4111	0.8950	0.8697	0.3013
ILMF-VH	0.4003	0.8933	0.9954	0.4111

CHERRY	0.3960	0.9112	0.9940	0.2354
PHISGAE	0.4587	0.9455	0.9959	0.4303

10. The statistical analysis for the comparisons of prediction performance.

Table S9. McNemar's Test results comparing our method with top three baseline methods.

Methods	McNemar's chi-squared	p-value	Significance(Y/N)
PHISGAE & NIMGSA	141.29	< 2.2e-16	Y
PHISGAE & ILMF-VH	7.3788	0.0066	Y
PHISGAE & CHERRY	85.923	< 2.2e-16	Y

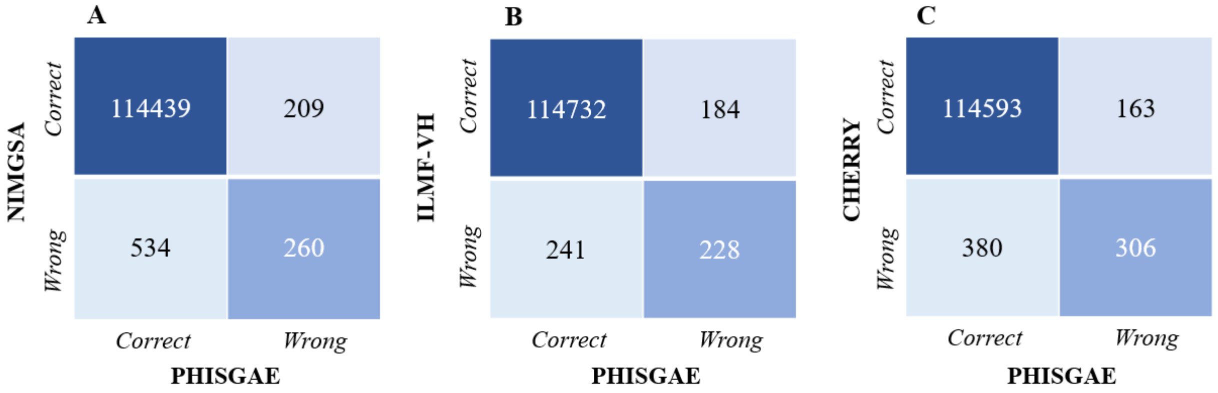


Figure S7. Confusion matrix for McNemar's Test. A): Confusion matrix derived from PHISGAE and NIMGSA. B): Confusion matrix derived from PHISGAE and ILMF-VH. C): Confusion matrix derived from PHISGAE and CHERRY.

Table S10. Paired t-test results comparing our method with top three baselines on important metrics.

Methods	Metrics	t-Stat	p-value	Significance(Y/N)
PHISGAE & NIMGSA	AUPR (+)	17.635	2.747e-08	Y
	AUC (+)	69.215	1.386e-13	Y
	ACC (-)	-1.8536	0.09679	N
	MCC (-)	-1.5316	0.16000	N
PHISGAE & ILMF-VH	AUPR (+)	16.792	4.218e-08	Y
	AUC (+)	59.311	5.548e-13	Y
	ACC (-)	-47.228	4.285e-12	Y
	MCC (-)	-1.3282	0.2168	N
PHISGAE & CHERRY	AUPR (+)	64.479	2.62e-13	Y
	AUC (+)	40.190	1.819e-11	Y
	ACC (+)	45.035	6.562e-12	Y
	MCC (+)	75.158	6.611e-14	Y

We performed paired sample t-tests on four evaluation metrics derived from the results of 10 repetitions of 5-fold cross-validation. We have confirmed that the data met the normality assumption using the Shapiro-Wilk test, which ensures the validity of our t-tests. The tests revealed statistically significant advantages in favor of our proposed method on metrics where it outperformed the baselines. Conversely, in instances where our method exhibited slightly lower performance, the statistical tests generally indicated non-significant differences. For example, although NIMGSA has slightly higher value on ACC and MCC, the P-value exceeds 0.05, indicating that the difference is not statistically significant. Furthermore, the relatively large standard deviation of NIMGSA's metrics suggests greater performance variability, likely due to the model's sensitivity to the data. Therefore, while NIMGSA performs better in certain metrics, its instability undermines the overall statistical significance of this advantage.

11. HTSR of Predicted Phages at Different Taxonomic Levels.

Table S11. HTSR of Predicted Phages at Different Taxonomic Levels.

Phage	Genus	Family	Order
NC_003387	100%	-	-
NC_017973	100%	-	-
NC_001900	100%	-	-
KJ000058	0	100%	-
HQ259103	40%	100%	-
JX442241	80%	80%	100%
NC_020416	20%	100%	-

12. The time complexity of PHISGAE.

The time complexity of our model is primarily influenced by several key components: 1) the GCN layers, where the graph convolution operations depend on the number of edges in the graph and the feature dimension; 2) the number of training epochs, which impacts the forward and backward passes during each iteration; and 3) the use of 5-fold cross-validation for model evaluation. The model also includes sparse matrix operations, data preprocessing, and metric calculations, but these have lower computational complexity compared to the GCN layers. Taking these factors into account, the overall time complexity of our model is $O(5TL|E|d^2)$, where T is the number of training epochs, L is the number of GCN layers, $|E|$ is the number of edges in the heterogeneous network, and d is the feature dimension of each node.