frontiers | Frontiers in Bioinformatics

TYPE Original Research
PUBLISHED 16 October 2025
DOI 10.3389/fbinf.2025.1622931

OPEN ACCESS

EDITED BY
Matthew Bashton,
Northumbria University, United Kingdom

REVIEWED BY
Yannan Bin,
Anhui University, China
Muhammad Shujaat,
Jeonbuk National University,
Republic of Korea

*CORRESPONDENCE
Narasaiah Kolliputi,
nkollipu@usf.edu

RECEIVED 05 May 2025
ACCEPTED 19 September 2025
PUBLISHED 16 October 2025

CITATION
Bodaka S and Kolliputi N (2025) CoMPHI: a novel composite machine learning approach utilizing multiple feature representation to predict hosts of bacteriophages.
Front. Bioinform. 5:1622931.
doi: 10.3389/fbinf.2025.1622931

CoMPHI: a novel composite machine learning approach utilizing multiple feature representation to predict hosts of bacteriophages

Shreyashi Bodaka[1] and Narasaiah Kolliputi[2]*

[1]Strawberry Crest IB High School, Dover, FL, United States, [2]Division of Allergy and Immunology, Department of Internal Medicine, Morsani College of Medicine, University of South Florida, Tampa, FL, United States

Phage therapy has reemerged as a compelling alternative to antibiotics in treating bacterial infections, especially for superbugs that have developed antibiotic resistance. The challenge in the broader application of phage therapy is identifying host targets for the vast array of uncharacterized phages obtained through next-generation sequencing. We introduce a Composite Model for Phage Host Interaction (CoMPHI) that integrates alignment-based approaches with machine learning. The model generates multiple feature encodings from nucleotide and protein sequences of both phages and hosts. It incorporates alignment scores between phage-phage, phage-host, and host-host pairs, creating a composite prediction framework. During 5-fold cross-validation, CoMPHI achieved Area Under the ROC Curve (AUC-ROC) values of 94–96.7% and accuracies of 92.3–95.1% across taxonomic levels from species to phylum. Comparative analysis showed a 6–8% performance improvement when alignment scores were included. Ablation studies demonstrated that combining nucleotide and protein encodings, along with phage-host, host-host, and phage-phage alignment scores, significantly enhanced prediction accuracy. CoMPHI provides a robust and comprehensive framework for predicting phage-host interactions. By combining sequence features and alignment information, the model advances computational tools that can accelerate the application of phage therapy in modern medicine.

KEYWORDS

sequence alignment, machine learning, bacteriophages, antibiotic resistance, phagehost prediction

1 Introduction

Antimicrobial resistance (AMR) was declared one of the top 10 global health threats by the World Health Organization (WHO). Antibiotics, considered a cornerstone of modern healthcare, are under threat from antibiotic resistance, which has emerged as a significant global public health and socioeconomic issue. An estimated 4.95 million deaths were attributed to bacterial antibiotic resistance, with 1.27 million deaths being specifically linked to bacterial AMR in 2019 (Murray et al., 2022). The World Bank projected that up to 3.8% of the global gross domestic product could be lost due to AMR by 2050 (Jonas et al., 2017). Among drug-resistant microbes, a significant threat is posed by the

group known as ESKAPEE, an acronym for *Enterococcus* faecium, *Staphylococcus aureus*, *Klebsiella pneumoniae*, *Acinetobacter* baumannii, *Pseudomonas aeruginosa*, *Enterobacter* spp., and *Escherichia coli*. These pathogens comprise high to critically drug-resistant strains and fall into the WHO's Critical Priority I and II categories. The pharmaceutical industry currently regards antibiotic development as financially imprudent (Safir et al., 2020) due to economic and regulatory barriers, resulting in diminished interest in this critical area (Bartlett et al., 2013). This circumstance heightens the imminent threat of entering a post-antibiotic era (Bartlett et al., 2013).

Phage therapy emerges as a compelling alternative to antibiotics in treating bacterial infections, particularly in combating superbugs that have developed resistance to traditional antibiotics. (Saw and Song, 2019). Phages are highly specific to the type of host they can infect (Saw and Song, 2019). This specificity implies that a particular phage would only target a particular strain or species of the host. Currently, culture-based or *in vitro* methods are employed to characterize and isolate phages that lyse their specific hosts. However, this method is resource-, labor-, and time-intensive, heavily dependent on the lytic cycle, limited to hosts that can be cultivated, not suited to be applied in large-scale or complex environments, and has low efficiency (Hyman, 2019). Leveraging state-of-the-art metagenomic sequencing and advanced bioinformatics, *in silico* prediction of putative hosts for metagenomic sequenced phages can accelerate and broaden the application of phage therapy in modern medicine. Such predictions are based on genomic signals arising from the coevolution and/or arms race between phages and hosts. These can be broadly categorized as alignment-based and alignment-free/machine learning methods.

Alignment-based methods leverage genomic and proteomic sequence homology/similarity to predict the host range of a phage (Versoza and Pfeifer, 2022). While this method can achieve high accuracy by controlling the alignment threshold, its recall is limited due to the constraints of sequences in the search database (Ahlgren et al., 2017). Additionally, it struggles to adapt to evolutionary shifts in phage and host genomes (Hall et al., 2013).

Alignment-free methods operate by extracting patterns and compositions from labeled empirical training data, employing statistical and/or probabilistic techniques, and do not rely on the alignment of sequences. The efficacy of a machine learning model hinges on capturing diverse genomic signals arising from intricate interactions between phages and hosts within a feature set (Li and Zhang, 2022). While numerous studies have explored various aspects of phage-host genomic signals as feature sets for machine learning models, the focus has primarily been on either phage nucleotide sequences or specific proteins, such as WIsH (Galiez et al., 2017). Remarkably, only a limited number of studies have integrated both. Additionally, very few studies have delved into feature sets derived from both phages and hosts, overlooking the evident close co-evolution of these entities. Furthermore, encoding genetic sequences using a specific method may highlight specific composition properties or patterns associated with their function, potentially overlooking functionalities due to data sparsity. Consequently, machine learning algorithms might miss features related to other functionalities within the sequences. This gap underscores the potential for enhanced model accuracy through a more comprehensive exploration of combined phage and host genomic features. Within these broader categories, alignment-based methods demonstrate high accuracy but low recall whereas alignment-free methods exhibit higher recall, but lower precision compared to alignment-based methods. *This paper introduces a novel composite model for predicting phage-host interactions, hypothesizing that capitalizing on the accuracy of alignment-based methods and the recall and flexibility of machine-learning techniques will improve its performance further than the current literature.* The model first utilizes multiple feature encodings from both nucleotide and protein sequences of phages and hosts. Then it leverages similarity scores from alignment-based methods for phage-phage, phage-host, and host-host interactions, along with a machine learning algorithm to predict the interaction probabilities between phages and hosts.

2 Materials and methods

2.1 Dataset collection and pre-processing

The dataset contained genomes of phages and hosts and phage-host interaction downloaded from the National Center of Biotechnology Information (NCBI) RefSeq bacteriophage database available as of August 2023 (US National Library of Medicine). This included 3,629 unique phage-host interactions between 3,629 phages and 815 hosts. Only phages that infect bacteria along with their hosts were selected and incomplete genomes of nucleotides and protein sequences were removed. To reduce bias due to the over-representation of a particular phage, data redundancy was removed by clustering phage genomes using CD-HIT (Fu et al., 2012) at a 95% identity match resulting in a dataset containing 3,018 unique phage-host interactions between 3,018 phages and 353 hosts. The 95% identity match was chosen to reduce redundancy and lower computational load without eliminating diversity of the dataset. The NCBI datasets tool was utilized to collect host taxonomy data from the phylum to genus level by inputting each host name into the tool at https://api.ncbi.nlm.nih.gov/datasets/v2alpha/taxonomy, which returned an Excel file with the host taxonomy data at each level. Then phage-host interactions with incomplete host taxonomy were removed from the data resulting in a final data set with 2,948 unique phage-host interactions between 2,948 phages and 256 hosts. As there is no laboratory-tested negative phage-host interaction data, a balanced set of negative interaction data was generated using phage-host interactions that are not included in this final cleansed dataset.

2.2 Composite model outline

The composite model comprises primarily three key components:

1. Generation of Alignment Bit Scores: This involves creating alignment bit scores for phage-phage, host-host, and phage-host interactions.
2. Generation of Multiple Feature Encodings: This step focuses on generating multiple feature encodings for the nucleotides and proteins of both phages and hosts.

3. Construction of a Composite Machine Learning Model: In this stage, a composite machine learning model is developed by combining alignment-free prediction using machine learning with the alignment-based bit scores.

2.3 Generation of alignment bit scores

Due to the process of co-evolution, phages and their hosts share common genetic elements. Consequently, there is a strong probability that a closely related host will be susceptible to the same phage, or that similar phages will target the same host. To leverage this close association, alignment bit scores were acquired for phage-host, host-host, and phage-phage interactions using NCBI BLAST (Altschul et al., 1990). Phage–phage alignment bit scores are acquired by establishing a reference phage database that includes all phages in the dataset. Each phage nucleotide in the dataset undergoes a search against this reference phage database using BLASTn with an e-value of 0.0001. The bit score of the maximum hit in this search (after excluding matches to itself) is documented as the phage–phage alignment score for the respective phage forming an array BIT_PP of dimension N, where N is the total number of unique phages in the dataset. Similarly, phage–host alignment bit scores are obtained by establishing a reference host database encompassing all hosts in the dataset. Each phage nucleotide in the dataset is then queried against this host reference database using BLASTn with an e-value of 0.0001. The bit score of the maximum hit in this search is recorded as the phage–host alignment score for the corresponding phage, resulting in an array BIT_PH of dimension NxM, where N is the total number of unique phages, and M is the total number of unique hosts in the dataset. Lastly, a search involving all hosts in the dataset against the reference host database, with an e-value of 0.0001, is conducted to register the host–host alignment scores. This process forms an array BIT_HH of dimension M, where M is the total number of unique hosts in the dataset.

2.4 Generation of multiple feature encodings

Utilizing multiple representations and a wider array of features extracted from nucleotides and proteins of both the phage and hosts enables harnessing complementary genetic signals from various levels of molecular interaction, contributing to an enhanced accuracy in phage-host prediction. To encode nucleotide sequences, feature encodings that are agnostic to the length of nucleotides were used to mitigate the influence of sequence length on bias. Following a similar approach as Li et al. (2021), feature encodings for protein sequences were generated using the iLearn tool (Chen et al., 2020). The list of these encodings can be found in Table 1.

As each phage or host consists of multiple protein sequences, six operators (mean, median, standard deviation, variance, maximum, and minimum) were employed to aggregate features derived from these multiple protein sequences. All feature encodings were normalized employing the min-max data normalization method, ensuring that the feature values fall within the range of 0–1. The feature encodings were consolidated into two feature vectors for each of the phage/host: one for nucleotide sequences and another for proteins.

To test deep learning models, the features were transformed into images. The individual sequential feature vectors originating from both phages and hosts underwent initial normalization employing the min-max data normalization method, ensuring that the feature values fall within the range of 0–1. These normalized vectors were then reshaped into an nxn array placing values into the array where n satisfies the condition: $(n-1) \times (n-1) < N$ and $N \leq n \times n$. In cases where $N \leq n \times n$, padding is applied by introducing zeros to the remaining $n \times n - N$ entries (Xu et al., 2020). A bilayer architectural framework, incorporating nucleotide and protein layers, was devised by stacking feature vectors derived from phages and hosts.

2.5 Construction of a composite machine learning model

Considering the exponential growth in genomic data and the objective of utilizing a single model for all viral genomes, a comparative study between several algorithms was conducted. Identifying key candidate algorithms was based on a review of the literature. Each algorithm differs based on underlying principles, assumptions, and approaches. These algorithms are also easy to implement and readily available via machine learning packages across multiple platforms/programming environments. Machine learning models that were evaluated are Logistic Regression (LR), Support Vector Machine (SVM), Multilayer Perceptron (MLP), Decisions Tree (DT), K-Nearest Neighbor (KNN), Random Forest (RF), and Convolution Neural Network (CNN). After evaluating the model performance, considering the principles of parsimony and computational complexity, RF was identified as the superior algorithm for predicting the putative host of a phage as shown in Figure 2. This analysis utilized feature vectors extracted from the nucleotide and protein sequences of both phages and hosts. Probabilistic classifier RFs were constructed with 100 trees, considering 1,004 features for possible splitting at each node. Next, variable importance was estimated using the impurity method to access the contribution percentages of each of the features.

To further enhance each of the evaluated machine learning model's performance, alignment-based bit scores between phage-phage, phage-host, and host-host were integrated with the machine learning prediction probabilities using a weighted sum as shown in the equation below. The host with the highest probability score after integrating the alignment bit scores was selected as the putative host of the phage.

$$\begin{aligned} Prfinal = Prm\,(1-r) + Pra\,(BlastPhage \\ - Host\,(pt,H)\,(1-a) \\ + BlastHost - Host\,(hs,H)a)\,r \end{aligned}$$

Prfinal is the prediction probability from the RF model computed for all hosts, with *Prm* denoting the prediction probability from the RF model and *Pra* denoting the prediction probability from the sequence alignments. r is the weighting between alignment-free and alignment-based methods, and a is the weighting between phage-host and host-host alignments.

- H is all hosts in the dataset.

TABLE 1 Details of nucleotide and protein features.

Type	Encoding	Formula	Details
DNA	Kmer	f(s)=(N(s))/N, s ∈ {AAA, AAC, AAG, …, TTT}; where s = 3	The occurrence frequencies of k neighboring nucleic acids where N(s) is the number of kmer type s, while N is the length of a nucleotide sequence
DNA	RCKmer	f(s)=(N(s))/N, s ∈ {AAA, AAC, AAG, …, TTT}; where s = 3	The occurrence frequencies of k neighboring nucleic acids where N(s) is the number of kmer type s, while N is the length of a nucleotide sequence, and reverse-complement kmers are removed
DNA	CKSNAP	(N_AA/N_total, N_AC/N_total…, N_TT/N_total)|k = 0	The frequency of nucleic acid pairs separated by any k nucleic acid, where k = 5
DNA	PseEIIP	V = [EIIP_AAA·f_AAA, EIIP_AAC·f_AAC, …,EIIP_TTT·f_TTT]	Mean EIIP values of trinucleotides in each sequence, f being the normalized frequency
DNA	NAC	f(t)=(N(t))/N, t ∈ {A, C, G, T(U)}	The frequency of each nucleic acid type in a nucleotide sequence, N being the length of the sequence
DNA	DNC	D(r,s) = N_rs/(N-1), r,s ∈ {A,C,G,T(U)}	Di-nucleotide composition, where N_rs is the number of di-nucleotides with nucleic acids r and s
DNA	TNC	D(r,s,t) = N_rst/(N-2), r,s,t ∈ {A,C,G,T(U)}	Tri-nucleotide composition, where N_rst is the number of tri-nucleotides with nucleic acids r, s, and t
Protein	MW	MW = sum(w_t)-(m-1)*18.01	The molecular weight of a protein sequence where w_t represents the molecular weight of the amino acid t
Protein	AC	AC = N_c, c∈{C,H,O,N,S}	The abundance of selected chemical elements composing a protein, where N_c is the number of occurrences of Carbon, Hydrogen, Oxygen, Nitrogen, and Sulfur in the sequence
Protein	AAC	N_t/N, t∈{A,C,D,E,F,G,H,I,K,L,M,N, P,Q,R,S,T,V,W,Y,*}	The frequency of each amino acid type in a protein or peptide sequence where t is the amino acid, N is the total number of amino acids

- *pt* represents the input/testing phage.
- *hs* represents the host of the most similar phage in the dataset based on BIT_PP
- *BlastPhage-Host* is using the BIT_PH scores between the phage and host (Figure 1E)
- *BlastHost-Host* is BIT_HH scores where the comparing host is the host of the top match from BIT_PP (Figures 1B–D)

2.6 Model optimization

Grid search was employed to fine-tune hyperparameters for the RF model. The optimal hyperparameter configuration identified through the grid search consisted of using 100 trees with a maximum depth of 20, a minimum number of samples required to split a node-set to 10, a minimum number of samples required at a leaf node set to 4, the maximum features set as sqrt, and a random state of 42. The parameters of the CNN can be found in the GitHub (linked at the end of the section).

Another iteration of the grid search was executed at 0.1 increments to determine the optimal weights for the contributions of machine learning predictions and alignment bit scores in the model. From this grid search, alpha and gamma values of 0.9 and 0.4, respectively, were identified. The code can be found at https://github.com/Bshrey/CoMPHI/tree/main.

3 Results

3.1 Model validation and performance

3.1.1 Performance of machine learning models

To measure performance, the following measures were used: accuracy (Acc), sensitivity (Sen), specificity (Spe), and area under the receiver-operating characteristic curve (AUC-ROC). 5-fold cross-validation was used on the entire dataset to ensure the generalization of the model and to assess model performance. Furthermore, to test the model on unseen data, testing was performed with the entire dataset using a randomized 70-30 split. This testing was repeated 10 times, and the averages were calculated to take care of data bias. Among all the algorithms included in the comparative analysis, RF and CNN demonstrated the best performance, at an accuracy and AUC-ROC of 86.5% and 88.5% respectively for RF, and 83.6%, and 88% respectively for CNN as illustrated in Figure 2. However, RF is a better algorithm due to its interpretability, computational simplicity, automatic feature importance, and simpler pre-processing of features. These metrics

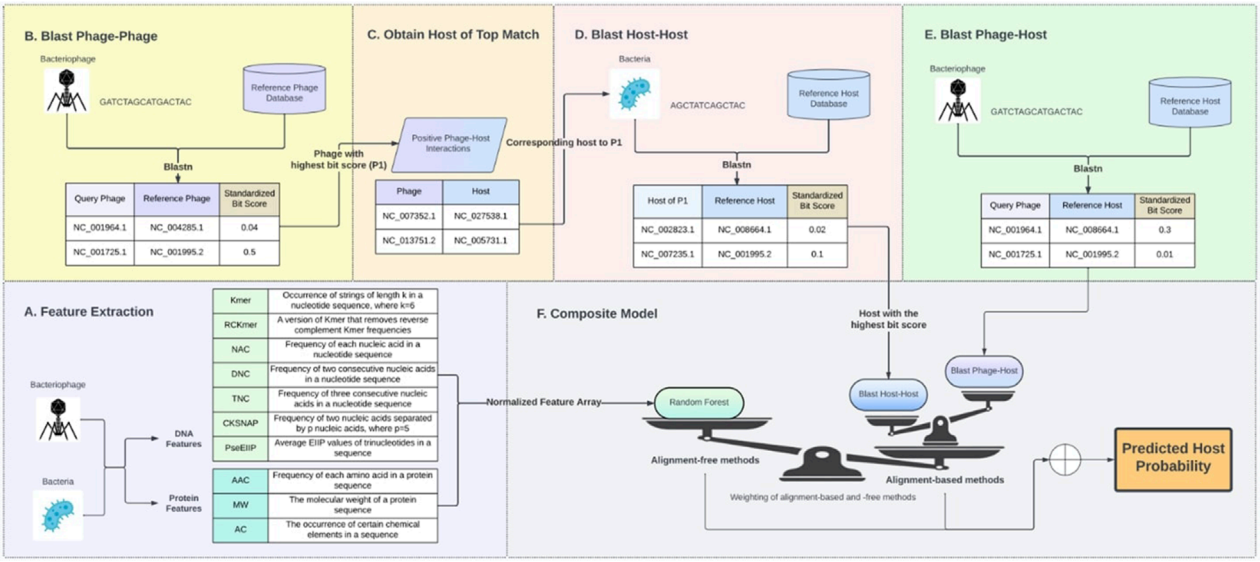

FIGURE 1
Composite model design. **(A)** Nucleotide and protein features from phage and host. **(B–E)** Alignment score matrices BIT_PP, BIT_HH, and BIT_PH. **(F)** Composite Model.

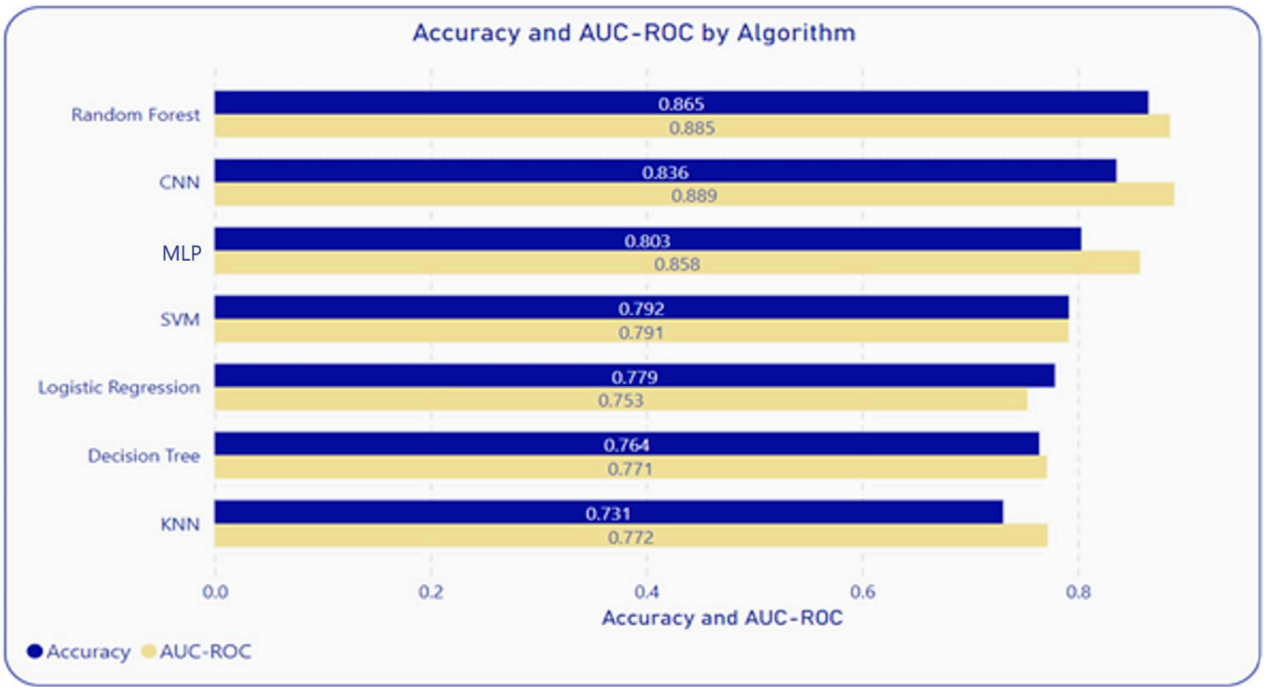

FIGURE 2
Performance comparisons of RF with other algorithms.

were within 2%–3% of the 5-fold cross-validation further proving the robustness of the RF model on unseen data.

3.1.2 Performance of composite models

The machine learning model's performance is improved by incorporating alignment scores. A random 70-30 split of the entire dataset was employed to assess the composite model. This testing process was iterated 10 times, and the averages were computed to address potential data bias. This resulted in AUC-ROC of 94.0%, 96.4%, 96.5%, 96.6%, 96.6%, and 96.7% and accuracy of 92.3%, 93.3%, 93.6%, 94%, 94.9%, and 95.1% at the Species, Genus, Family, Order, Class, and Phylum levels, respectively. These taxonomy

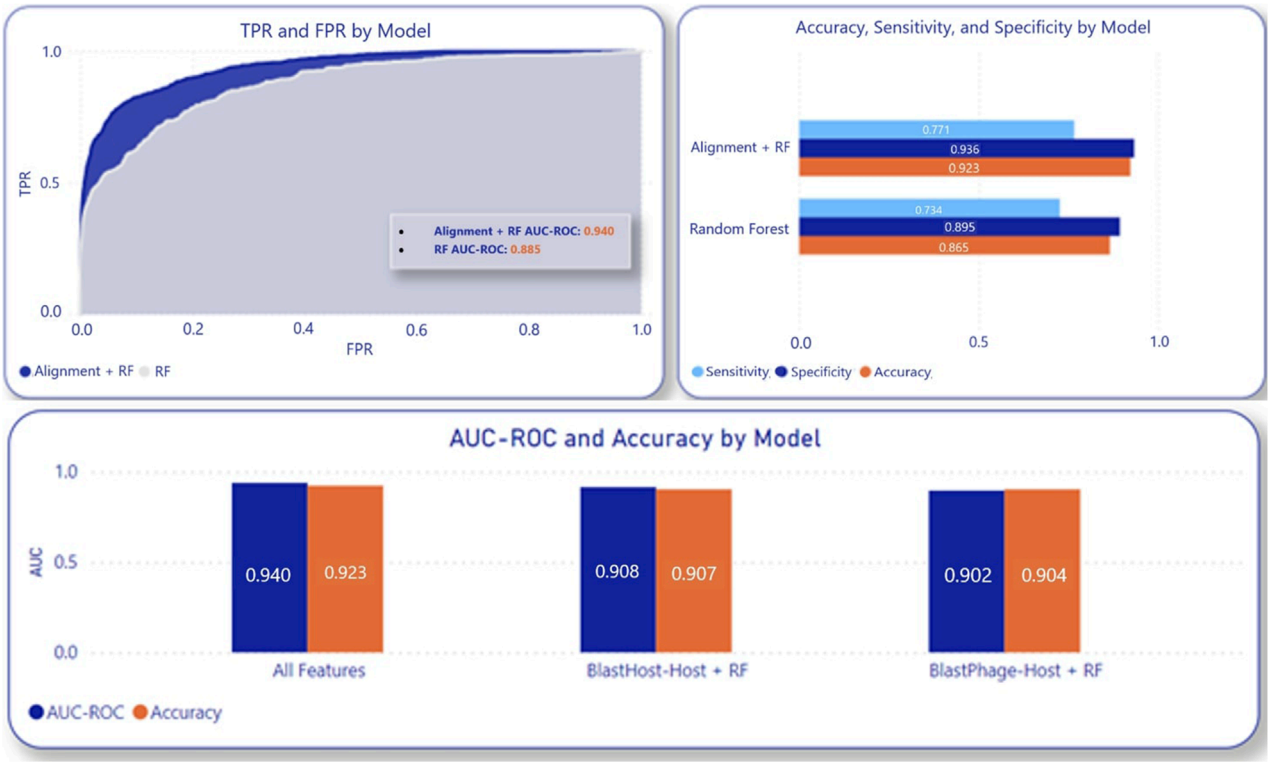

FIGURE 3
Composite model performance with inclusion of alignment score components compared to RF alone.

level metrics were determined by using the previously obtained host taxonomy and having the model predict the corresponding taxonomic level in addition to the species of the host. When compared to the model utilizing only the machine learning algorithm, the composite model demonstrates approximately a 6-point higher AUC-ROC, as well as improved sensitivity, specificity, and accuracy, as illustrated in Figure 3. 5- fold cross validation using the entire dataset on the composite model also led to accuracies in 2%–3% of 70–30 testing. The sensitivity and specificity were also improved by adding alignment-based scores, as seen in Figure 3, showing that alignment-based scores aid in improving True Positive Rate and True Negative Rate.

3.2 Ablation study

To understand the contributions of each of the components in the composite model with multiple features, an ablation analysis was conducted to compare the impact of including or excluding features. The models were evaluated based on the following feature combinations: only nucleotide features from phages, only protein features from phages, only nucleotide features from hosts, only nucleotide features from both phages and hosts, only protein features from both phages and hosts, and a combination of nucleotide and protein features from both phages and hosts. This analysis led to the conclusion that utilizing both nucleotide and protein features from both phages and hosts resulted in the highest prediction accuracies, as depicted in Figure 4.

Furthermore, nucleotide features had a greater influence on the prediction than proteins. Host nucleotides and/or proteins contributed more to the model performance than the phage nucleotides and/or proteins. Another ablation analysis was conducted to assess the impact of individual alignment scores—phage-phage, phage-host, and host-host—on the composite model. In this analysis, the following combinations were tested using 70-30 randomized data split on the entire dataset. The test was repeated 10 times, and scores were averaged across these tests: BlastPhage-Host with the RF model, BlastHost-Host with the RF model, and BlastHost-Host with BlastPhage-Host and the RF model. As shown in Figure 3, the composite model using BlastHost-Host with BlastPhage-Host and the RF model scored the highest, indicating that utilizing all alignment scores helps increase accuracy.

4 Discussion

4.1 Variable importance

As proposed in this paper and proven by the ablation studies, composite features from both alignment-based and alignment-free methods as well as including multiple feature encodings from both nucleotides and proteins of phage and host have significantly improved the model's performance in predicting phage-host interactions. To further gain a deeper understanding of the contribution of each alignment-free feature encoding, variable importance was assessed using the impurity method in the RF

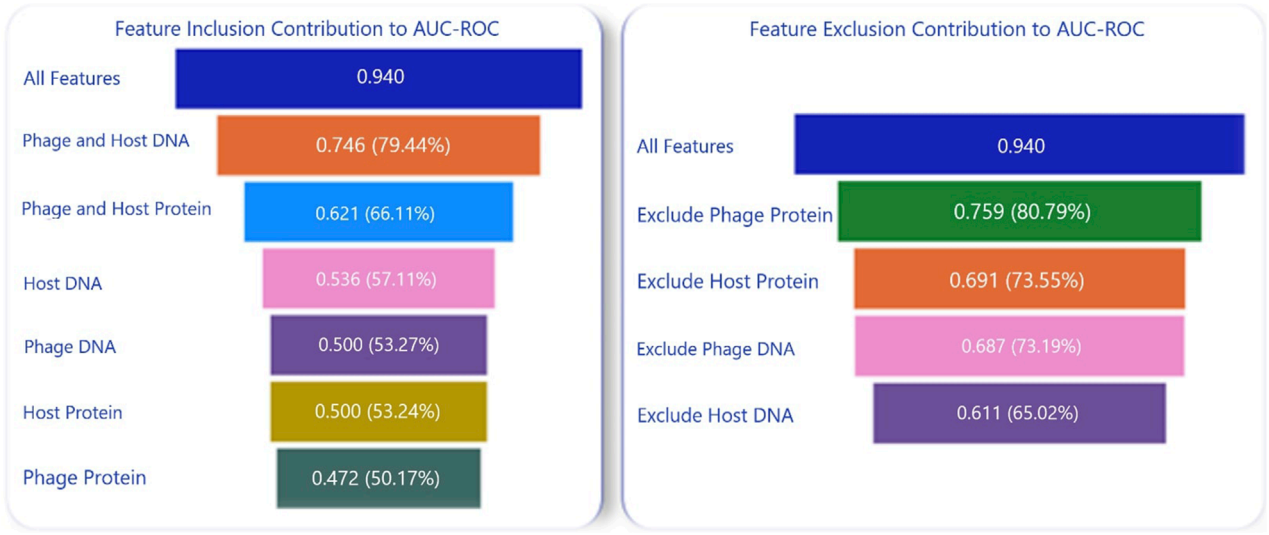

FIGURE 4
Composite Model Performance with Inclusion and Exclusion of Nucleotide and Protein Features. The parenthetical values are the percentages of the original AUC-ROC that each ablation retains.

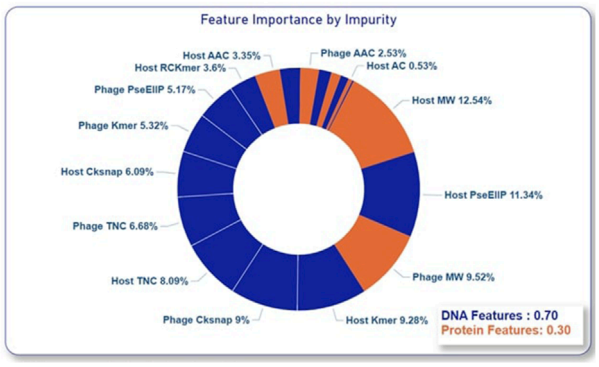

FIGURE 5
Variable importance estimation using impurity method.

model. This estimation further proved the results of the ablation study that the nucleotide features made a greater contribution to the prediction accuracy compared to the protein features as shown in Figure 5.

4.2 ESKAPEE testing

To assess the prediction sensitivity of the model concerning different groups of hosts within a taxonomic group, further testing was conducted. This involved the following procedures:

1. Exclusion of prevalent ESKAPEE host families from the training set and subsequent testing of prediction accuracies.
2. Inclusion of only prevalent ESKAPEE host families from the training set and subsequent testing of prediction accuracies.

Results from both tests revealed no drastic changes in the prediction accuracies of the composite RF model. This suggests that the model retains its predictive capability and can be effectively employed to predict phages for currently prevalent pathogens. It also suggests that the model retains its performance for hosts that are not common in the current dataset proving that the model will reliably predict phages for any novel hosts that are discovered.

4.3 Improvement

One significant benefit of the composite RF model lies in its versatility, allowing for straightforward expansion to include additional meaningful features that can enhance our understanding of phage–host interactions in future studies. All features following the phage infection cycle such as CRISPR spacers and auxiliary metabolic genes or tRNAs can be included in the feature vectors. This paper only includes feature encodings that are not dependent on the sequence length. Further studies could expand this model to not be restricted by the sequence length.

5 Conclusion

Phage therapy is already being used in the personalized treatment of patients for whom traditional antibiotics have failed to work (Yang et al., 2023). Culture-based methods of identifying the host range of a phage are time and labor-intensive and hence can be a bottleneck in the widespread use of phage therapy in modern medicine, especially with the exponential increase in phage classifications by next-gen sequencing methodologies (Klumpp et al., 2012). Recent advancements in computational and bioinformatics tools have made it possible to predict a putative host for a phage with high accuracy, thus

reducing the time and effort required to experimentally test a phage's host range. This paper introduces a novel composite machine learning model that leverages alignment-free methods, by incorporating multiple feature encodings from both nucleotide and protein sequences of phages and hosts and combines it with alignment-based features of alignment scores between phage-phage, phage-host, and host-host. The composite model is not only robust as proven by the 5-fold validation and 70-30 testing but is also interpretable as proven by the ablation studies and variable importance analysis (Li and Zhang, 2022). By incorporating alignment-based scores alongside multiple features from phage and host, the model achieves a notable 5%–6% increase in accuracy and AUC-ROC. Ablation analysis and variable importance analysis illustrate that nucleotide features contribute more to the performance than proteins, host nucleotides, and proteins have a greater influence than that of phages, and all alignment scores have an equal influence on the performance gain. These results indicate that the composite machine learning model is a promising solution in predicting phage-host interaction. This model can also be used for other classification problems involving nucleotide and protein sequences.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

SB: Writing – original draft, Investigation, Writing – review and editing, Methodology, Formal Analysis, Conceptualization, Validation, Visualization, Data curation. NK: Writing – original draft, Supervision, Writing – review and editing.

Funding

The author(s) declare that no financial support was received for the research and/or publication of this article.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The author(s) declared that they were an editorial board member of Frontiers, at the time of submission. This had no impact on the peer review process and the final decision.

Generative AI statement

The author(s) declare that no Generative AI was used in the creation of this manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher's note

References

Ahlgren, N. A., Ren, J., Lu, Y. Y., Fuhrman, J. A., and Sun, F. (2017). Alignment-free d_2^{*} oligonucleotide frequency dissimilarity measure improves prediction of hosts from metagenomically-derived viral sequences. *Nucleic Acids Res.* 45 (1), 39–53. doi:10.1093/nar/gkw1002

Altschul, S., Gish, W., Miller, W., Myers, E., and Lipman, D. (1990). Basic local alignment search tool. *J. Mol. Biol.* 215, 403–410. doi:10.1016/S0022-2836(05)80360-2

Bartlett, J. G., Gilbert, D. N., and Spellberg, B. (2013). Seven ways to preserve the miracle of antibiotics. *Clin. Infect. Dis.* 56 (10), 1445–1450. doi:10.1093/cid/cit070

Chen, Z., Zhao, P., Li, F., Marquez-Lago, T. T., Leier, A., Revote, J., et al. (2020). iLearn: an integrated platform and meta-learner for feature engineering, machine-learning analysis and modeling of DNA, RNA and protein sequence data. *Briefings Bioinforma.* 21 (3), 1047–1057. doi:10.1093/bib/bbz041

Fu, L., Niu, B., Zhu, Z., Wu, S., and Li, W. (2012). CD-HIT: accelerated for clustering the next generation sequencing data. *Bioinformatics* 28 (23), 3150–3152. doi:10.1093/bioinformatics/bts565

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 (19), 3113–3114. doi:10.1093/bioinformatics/btx383

Hall, J. P., Harrison, E., and Brockhurst, M. A. (2013). Viral host-adaptation: insights from evolution experiments with phages. *Curr. Opin. Virology* 3 (5), 572–577. doi:10.1016/j.coviro.2013.07.001

Hyman, P. (2019). Phages for phage therapy: isolation, characterization, and host range breadth. *Pharmaceuticals* 12 (1), 35. doi:10.3390/ph12010035

Jonas, O. B., Irwin, A., Berthe, F. C. J., Le Gall, F. G., and Marquez, P. V. (2017). Drug-resistant infections: a threat to our economic future (vol. 2): final report. HNP/agriculture global antimicrobial resistance initiative.

Klumpp, J., Fouts, D. E., and Sozhamannan, S. (2012). Next generation sequencing technologies and the changing landscape of phage genomics. *Bacteriophage* 2 (3), 190–199. doi:10.4161/bact.22111

Li, M., Wang, Y., Li, F., Zhao, Y., Liu, M., Zhang, S., et al. (2021). A deep learning-based method for identification of bacteriophage-host interaction. *IEEE/ACM Trans. Comput. Biol. Bioinforma.* 18 (5), 1801–1810. doi:10.1109/tcbb.2020.3017386

Li, M., and Zhang, W. (2022). PHIAF: prediction of phage-host interactions with GAN-based data augmentation and sequence-based feature fusion. *Briefings in Bioinformatics* 23 (1), bbab348. doi:10.1093/bib/bbab348

Murray, C. J., Ikuta, K. S., Sharara, F., Swetschinski, L., Aguilar, G. R., Gray, A., et al. (2022). Global burden of bacterial antimicrobial resistance in 2019: a systematic analysis. *Lancet* 399 (10325), 629–655. doi:10.1016/s0140-6736(21)02724-0

Safir, M. C., Bhavnani, S. M., Slover, C. M., Ambrose, P. G., and Rubino, C. M. (2020). Antibacterial drug development: a new approach is needed for the field to survive and thrive. *Antibiotics* 9 (7), 412. doi:10.3390/antibiotics9070412

Saw, P. E., and Song, E. W. (2019). Phage display screening of therapeutic peptide for cancer targeting and therapy. *Protein and Cell* 10 (11), 787–807. doi:10.1007/s13238-019-0639-7

Versoza, C. J., and Pfeifer, S. P. (2022). Computational prediction of bacteriophage host ranges. *Microorganisms* 10 (1), 149. doi:10.3390/microorganisms10010149

Xu, Y., Zhang, Z., You, L., Liu, J., Fan, Z., and Zhou, X. (2020). scIGANs: single-cell RNA-Seq imputation using generative adversarial networks. *Nucleic acids Res.* 48 (15), e85. doi:10.1093/nar/gkaa506

Yang, Q., Le, S., Zhu, T., and Wu, N. (2023). Regulations of phage therapy across the world. *Front. Microbiol.* 14, 1250848. doi:10.3389/fmicb.2023.1250848