PLOS ONE

RESEARCH ARTICLE

Protein embeddings improve phage-host interaction prediction

Mark Edward M. Gonzales[1,3], **Jennifer C. Ureta**[1,3], **Anish M. S. Shrestha**[1,2,3]*

1 Bioinformatics Laboratory, Advanced Research Institute for Informatics, Computing and Networking, De La Salle University, Manila, Philippines, **2** Systems and Computational Biology Research Unit, Center for Natural Sciences and Environmental Research, De La Salle University, Manila, Philippines, **3** Department of Software Technology, College of Computer Studies, De La Salle University, Manila, Philippines

* anish.shrestha@dlsu.edu.ph

Check for updates

OPEN ACCESS

Citation: Gonzales MEM, Ureta JC, Shrestha AMS (2023) Protein embeddings improve phage-host interaction prediction. PLoS ONE 18(7): e0289030. https://doi.org/10.1371/journal.pone.0289030

Editor: Iddya Karunasagar, Nitte University, INDIA

Received: March 21, 2023

Accepted: July 7, 2023

Published: July 24, 2023

Copyright: © 2023 Gonzales et al. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: All relevant data and scripts can be found at https://github.com/bioinfodlsu/phage-host-prediction.

Funding: This research was partly funded by the Department of Science and Technology Philippine Council for Health Research and Development (DOST-PCHRD) under the e-Asia JRP 2021 Alternative therapeutics to tackle AMR pathogens (ATTACK-AMR) program. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Abstract

With the growing interest in using phages to combat antimicrobial resistance, computational methods for predicting phage-host interactions have been explored to help shortlist candidate phages. Most existing models consider entire proteomes and rely on manual feature engineering, which poses difficulty in selecting the most informative sequence properties to serve as input to the model. In this paper, we framed phage-host interaction prediction as a multiclass classification problem that takes as input the embeddings of a phage's receptor-binding proteins, which are known to be the key machinery for host recognition, and predicts the host genus. We explored different protein language models to automatically encode these protein sequences into dense embeddings without the need for additional alignment or structural information. We show that the use of embeddings of receptor-binding proteins presents improvements over handcrafted genomic and protein sequence features. The highest performance was obtained using the transformer-based protein language model ProtT5, resulting in a 3% to 4% increase in weighted F1 and recall scores across different prediction confidence thresholds, compared to using selected handcrafted sequence features.

Introduction

One of the most pressing threats to global health is antimicrobial resistance (AMR), a phenomenon wherein microorganisms evolve to withstand exposure to bacteriostatic and bactericidal drugs. In 2019, 4.95 million AMR-related and 1.27 million AMR-attributable deaths were estimated [1]. In developing countries, this problem is compounded by the unregulated dispensation of antibiotics as a form of self-medication even for mild conditions [2, 3] and their routine use in the agricultural sector for disease prophylaxis [4] and livestock growth promotion [5].

A solution that is actively being explored to combat this problem is phage therapy, which capitalizes on the specificity of bacteriophages (hereinafter referred to as *phages*) to a narrow range of hosts. Phages have been shown to antagonize the target bacteria with minimal side effects and without triggering a dysbiosis of the beneficial microbiota [6]. However, the

foremost challenge to formulating phage cocktails for treating bacterial infections is identifying putative phages that attack the offending pathogens. Aside from being time- and cost-intensive, *in vitro* experiments require the cultivation of microbes under strict laboratory conditions, posing a bottleneck to the rapid selection of candidate phages.

With the advent of high-throughput sequencing technologies and the resulting increase in omic data, *in silico* approaches have been employed to help shortlist candidate phages. These can be broadly categorized into alignment-based methods [7, 8], which rely on sequence similarity to infer phage-host pairs, and alignment-free methods [9–11], which exploit features related to sequence composition, such as oligonucleotide frequency and codon usage bias. These reflect shared genomic properties that arise from the close coexistence and coevolution of phages and their hosts [12, 13].

Machine learning algorithms for phage-host interaction prediction have also been actively explored. Feature sets extracted from protein sequences include molecular weight [14], aromaticity [15], amino acid composition [14], protein-protein and domain-domain interaction [14, 16], and protein secondary structure [15]. Meanwhile, those obtained from genomic sequences include *k*-mer frequency [14, 15, 17], guanine-cytosine content [15], codon usage bias [15, 16], oligonucleotide frequency [16, 18], and shared transfer ribonucleic acids [19]. Recent studies have also investigated the application of deep learning architectures, primarily convolutional neural networks, that take these handcrafted properties as input [20–23].

While these existing models have been successful in integrating various features to improve their performance, most consider the entire proteome of both the phages and their hosts [14, 16, 18–21, 23], when only specific proteins are actually involved in phage-host interaction [24]. To initiate infection, a tailed phage typically adsorbs to the host bacterium's surface through receptor-binding proteins (RBPs) located at its tail's distal end [25]. In this regard, RBPs (e.g., tail fibers and spikes) serve as key machinery for host recognition and specificity [26–29].

Moreover, most tools for phage-host interaction prediction rely on manual feature engineering to transform raw sequences into numerical vectors, often requiring additional alignment or structural information [30–32]. The multitude of potentially informative signals that can be derived from these sequences also poses difficulty in selecting which features should be fed to the model [30].

Representation learning, in which raw biological sequences are converted into dense vectors in high-dimensional space, has recently been applied to some prototypical bioinformatics tasks, such as predicting protein function [33], succinylation sites [34], and sequence conservation [35]. The embeddings produced by protein language models, such as SeqVec [36] (which adopts the architecture of the natural language model ELMo [37]) and the transformer-based Evolutionary Scale Modeling (ESM) [38] and ProtTrans [31], have been demonstrated to capture protein secondary structure and physicochemical characteristics, features that are relevant to phage-host specificity [31, 38]. Representation learning also serves as a promising approach to address some of the limitations posed by the problem of data scarcity in phage-host datasets [13], as knowledge is distilled from large-scale protein databases [39–42]. However, the application of protein embeddings to the problem of phage-host interaction prediction remains unexplored.

In an attempt to address these gaps, our study seeks to contribute the following:

- We framed phage-host interaction prediction as a multiclass classification problem, with the embedding of a receptor-binding protein (RBP) as the input and the host genus as the output.

- We extensively tested different protein language models to automatically generate dense embeddings of RBP sequences.
- We constructed a random forest model for predicting phage-host interaction and showed that embeddings of RBPs outperform handcrafted genomic and protein sequence features, with the use of the protein language model ProtT5 resulting in the highest performance.

Materials and methods

In this section, we discuss the methodology of our study (Fig 1). For the purpose of reproducibility, the data and source code for all our experiments and analysis are publicly available at https://github.com/bioinfodlsu/phage-host-prediction. Detailed instructions on setting up the environment, installing the required dependencies, and running the code are also provided in this repository.

Data collection and preprocessing

We collected genome sequences of 18,389 phages, along with their proteome sequences (whenever available), via INPHARED [46], a pipeline for retrieving phage sequences from GenBank [47]; the sequences were downloaded in September 2022, and proteomes were retrieved for 16,836 phages. Limiting the host information to the genus level, we subjected the entries to preprocessing of host information and selection of annotated receptor-binding proteins, which we describe in detail in the following subsections.

Preprocessing of host information. INPHARED [46] returns host data for 15,739 phages across 278 different host genera. For entries where the host name is unspecified, we referred to the isolation host information in GenBank whenever available and filtered out nonbacterial isolation hosts. We used Biopython [48], an open-source suite of bioinformatics tools written in Python, in parsing the phages' GenBank records. We then standardized the host names following NCBI Taxonomy's nomenclature [49]. Note that, while some phages are known to be polyvalent (multihost), only five phage entries were recorded with multiple host genera. Hence, for simplicity, we mapped each polyvalent phage to its host with the highest number of interacting phages in the dataset.

After preprocessing, host information was supplied for an additional 84 phages, thus totaling 15,823 phages across 279 host genera; the additional identified genus, *Silvanigrella*, was from the isolation host of MWH-Nonnen-W8red.

Identification of receptor-binding proteins. Among the phage entries with host data, 15,158 entries have gene annotations, while the remaining 665 do not. For those with annotations, we selected the annotated receptor-binding proteins (RBPs) using a regular expression (S1 Listing in S1 File) and a manual exclusion list adapted from Boeckaerts *et al.* [50]; this exclusion list covers proteins related to RBPs but are not RBPs themselves (e.g., assembly and portal proteins). We also discarded sequences with undetermined amino acids (X).

Meanwhile, we ran the genomes of phages without annotation through Prokka [51], a wrapper tool for genome annotation. It first calls Prodigal [52] to predict the coding sequence coordinates. To identify the putative gene products, we configured Prokka [51] to refer to PHROG [53], a database of viral protein family clusters generated by employing hidden Markov model profile-profile comparisons for remote homology detection. PHROG [53] has also been used in previous studies that require the functional annotation of phage protein sequences [54, 55]. The annotated RBPs were selected following the same scheme described in the previous paragraph.

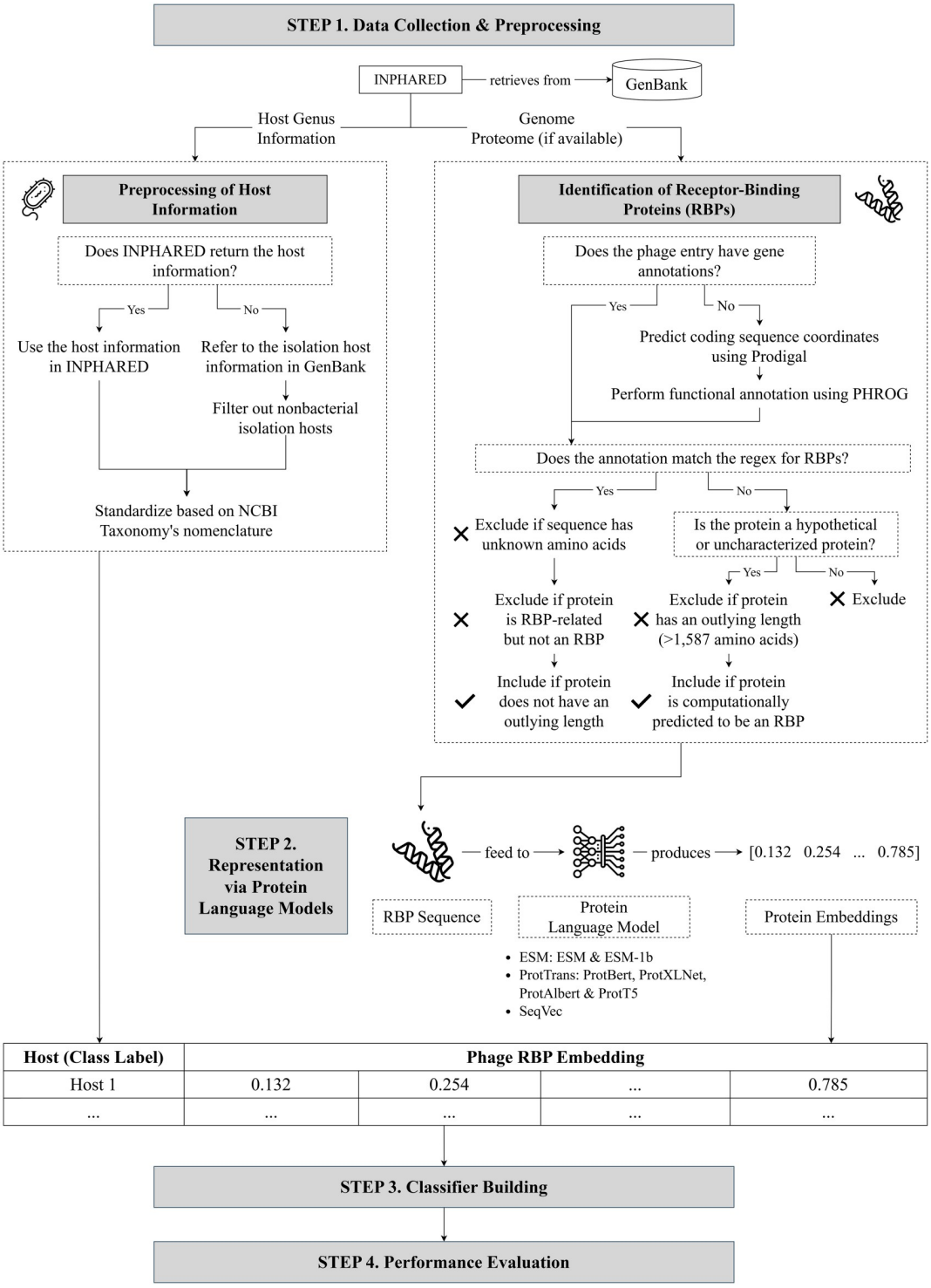

Fig 1. Methodology of our study. Step 1: We collected phage genomes, along with their proteomes (see Section Data collection and preprocessing for more details), and performed preprocessing to obtain the host information (see Section Preprocessing of host information) and select annotated receptor-binding proteins or RBPs (see Section Identification of receptor-binding proteins). Step 2: We fed the RBP sequences to pretrained protein language models to generate meaningful dense embeddings (see Section Representation via protein language models). Step 3: We built a random forest model with the RBP embeddings as the input and the host genus as the predicted output (see Section Classifier building). Step 4: We evaluated our model's performance (see Section Performance evaluation). Flat icons used in this figure are taken from [43–45].

https://doi.org/10.1371/journal.pone.0289030.g001

Afterwards, we discarded RBPs with lengths outside the interval $[Q_1 - 1.5 \cdot IQR, Q_3 + 1.5 \cdot IQR]$, where Q_1 is the first quartile, Q_3 is the third quartile, and IQR is the interquartile range of the RBP lengths. This resulted in the removal of protein sequences longer than 1,587 amino acids. Fig 2A and 2B show the distribution of the lengths of the RBPs before and after this step, respectively.

Finally, to expand the list of RBPs in our dataset, we also considered proteins labeled as hypothetical by GenBank or uncharacterized by Prokka. Following Boeckaerts *et al.* [50], we encoded these sequences via the transformer-based protein language model ProtBert [31] and fed the generated embeddings to their proposed extreme gradient boosting model to computationally predict whether the hypothetical proteins are RBPs or not. In total, our dataset consists

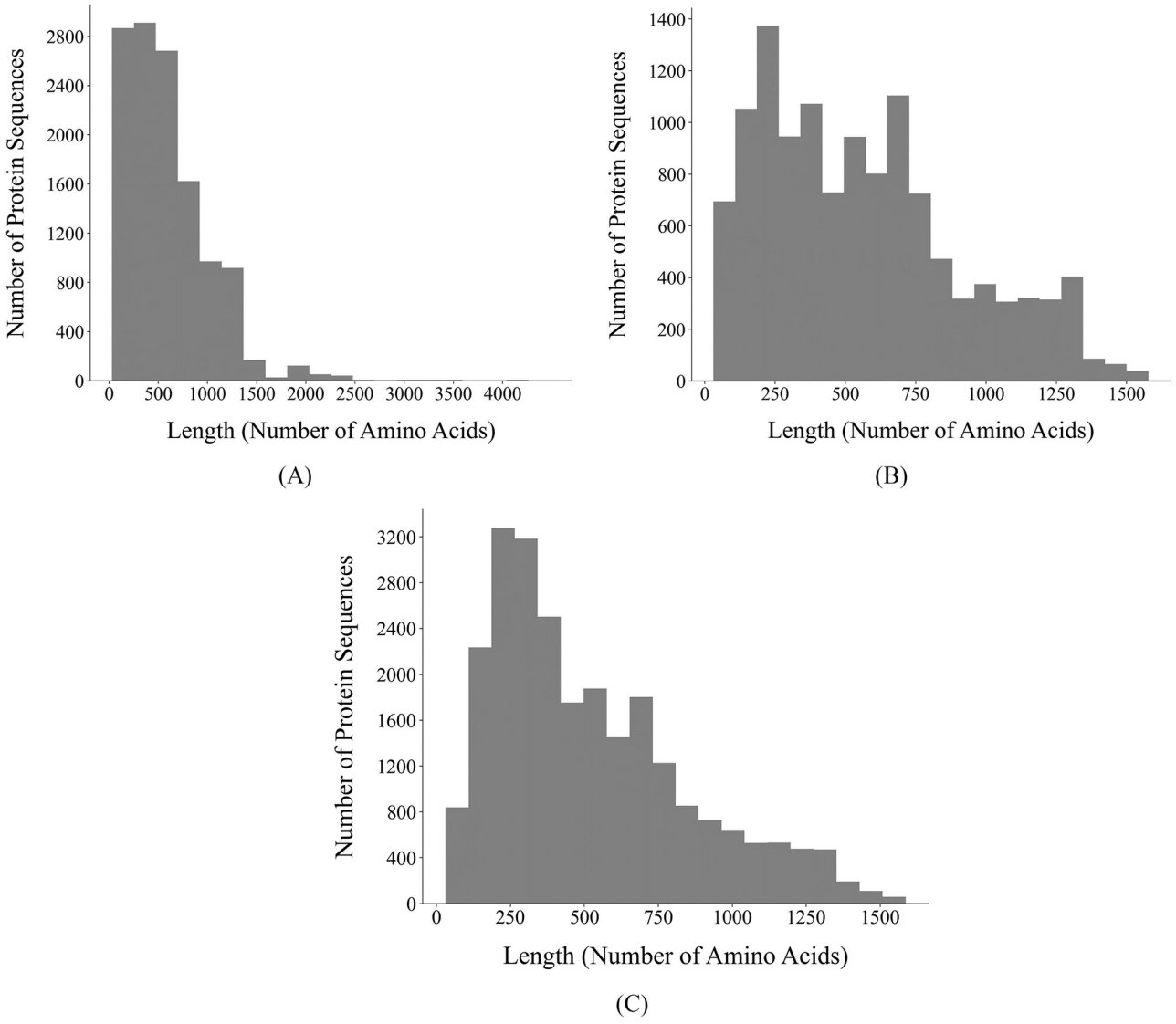

Fig 2. Distribution of the lengths of the receptor-binding proteins (RBPs). (A) Distribution of the lengths of the annotated RBPs selected based on annotation in GenBank and the functional annotation obtained using PHROG [53]. (B) Distribution of the lengths of the annotated RBPs after excluding those longer than 1,587 amino acids. This cutoff was set by defining outlying lengths as those outside the interval $[Q_1 - 1.5 \cdot IQR, Q_3 + 1.5 \cdot IQR]$, where Q_1 is the first quartile, Q_3 is the third quartile, and IQR is the interquartile range of the RBP lengths. (C) Distribution of the lengths of all the RBPs in our dataset, including those computationally predicted via the approach proposed by Boeckaerts *et al.* [50].

https://doi.org/10.1371/journal.pone.0289030.g002

Table 1. Statistics on the identification of receptor-binding proteins (RBPs). GenBank refers to the RBPs annotated in GenBank, PHROG refers to those selected based on the functional annotation obtained using PHROG [53]. *Predicted* refers to those computationally predicted via the approach proposed by Boeckaerts *et al.* [50].

	Num. of RBPs	Num. of Phages	Num. of Hosts
GenBank	11,133	5,805	167
PHROG	1,010	419	30
Predicted	12,609	5,941	192
Total	24,752	9,583	232

https://doi.org/10.1371/journal.pone.0289030.t001

of 24,752 RBPs across 9,583 phages and 232 hosts (Table 1); the distribution of the lengths of these RBPs is presented in Fig 2C.

Representation via protein language models

To generate dense vector representations (embeddings) of the RBP sequences, we explored seven pretrained protein language models as feature encoders: ESM [38], ESM-1b [38], ProtBert [31], ProtXLNet [31], ProtAlbert [31], ProtT5 [31], and SeqVec [36]. To this end, we used `bio_embeddings` [56], an open-source Python library that provides reproducible workflows for applying representation learning to protein sequences. These protein language models adopt deep learning architectures and are pretrained on large-scale protein databases with the goal of capturing biophysical features that may be relevant to downstream tasks.

SeqVec [36] adopts ELMo [37], which consists of a character-level convolutional neural network and two bidirectional long short-term memory networks. ESM1-b [38] is a 33-layer transformer built by optimizing the hyperparameters of the 34-layer transformer ESM. [38]. ProtBert [31], ProtAlbert [31], and ProtT5 [31] are 30-layer, 12-layer, and 24-layer autoencoding transformers, respectively. ProtXLNet [31] is a 30-layer autoregressive transformer. Additional technical details about these protein language models are provided in Table 2.

Since these protein language models output embeddings per "token" (i.e., per residue), we performed averaging over the residues to produce fixed-length protein embeddings [31, 36, 38]. Formally, suppose a protein sequence with r residues is fed as input to a protein language model that encodes each residue as an embedding of length s. The output of this model is an $r \times s$ matrix $\mathbf{M}$ whose i^{th} row is the embedding of the i^{th} residue. From this matrix $\mathbf{M}$, we obtain

Table 2. Protein language models for generating receptor-binding protein embeddings.

Model	Architecture	Pretraining Datasets	Vector length
SeqVec [36]	ELMo [37]	UniRef50 [40]	1024
ESM [38]	Transformer	UniParc [39]	1280
ESM-1b [38]	Transformer	UniRef50 [40]	1280
ProtBert [31]	Transformer	UniRef100 [40], BFD100 [41, 42]	1024
ProtXLNet [31]	Transformer	UniRef100 [40]	1024
ProtAlbert [31]	Transformer	UniRef100 [40]	4096
ProtT5 [31]	Transformer	UniRef50 [40], BFD100 [41, 42]	1024

https://doi.org/10.1371/journal.pone.0289030.t002

the protein embedding of length s, which we denote by $\mathbf{v}$, via Eq (1).

$$\mathbf{v} = \left[\frac{1}{r}\sum_{i=1}^{r}\mathbf{M}_{i1},\ \frac{1}{r}\sum_{i=1}^{r}\mathbf{M}_{i2},\ \frac{1}{r}\sum_{i=1}^{r}\mathbf{M}_{i3},\ \cdots,\ \frac{1}{r}\sum_{i=1}^{r}\mathbf{M}_{is}\right] \quad (1)$$

Note that, for ESM and ESM-1b (which accept sequences with at most 1022 residues only), we followed the approach of Marquet *et al*. [35] and split sequences longer than 1022 residues into non-overlapping subsequences of length 1022 (whenever possible). We fed each subsequence to the protein language model separately, resulting in a set of matrices corresponding to the per-residue embeddings. We then stacked these matrices into an $r \times s$ matrix $\mathbf{M}$ as described in the previous paragraph, and applied Eq (1) to obtain the protein embedding.

Classifier building

We framed phage-host interaction prediction as a multiclass classification problem, with the protein embeddings of the RBPs as the input and the host (at the genus level) as the output. Including all 232 hosts in our dataset resulted in class imbalance, with a quarter of the hosts already accounting for 96.02% of the dataset entries. To mitigate this, we restricted the class labels to only the top 25% (i.e., 58) hosts associated with the most RBPs. The class labels are enumerated in S1 Table in S1 File.

We divided our dataset into two sets D_1 and D_2, where D_1 contains the RBPs with class labels belonging to the top 25% hosts and D_2 contains the remaining entries. We then partitioned D_1 following a stratified 70%-30% train-test split and appended D_2 to the test set. As such, our test set includes RBPs with class labels outside the top 25% hosts, which we will refer to as the *others* class. This is to make the evaluation more reflective of real-world use cases, where we might encounter inputs not belonging to any class for which the model was trained. In total, our training and test sets have 16,636 and 8,116 samples, respectively. Table 3 reports the training and test set statistics on the top 10 hosts associated with the most RBPs; the complete statistics are given in S1 Table in S1 File.

Afterwards, we fed the protein embeddings to a random forest classifier built using `scikit-learn` [57], an open-source Python library for machine learning. In a further attempt to address class imbalance, we employed a weighted random forest model that

Table 3. Number of training and test samples for the top 10 hosts associated with the most receptor-binding proteins (RBPs). The RBPs associated with the top 10 hosts comprise 64.66% of our dataset. Four of the top 10 hosts (*Escherichia*, *Salmonella*, *Klebsiella*, and *Erwinia*) belong to the same order: Enterobacterales.

Host	Training Set	Test Set	Total
Escherichia	3,021	1,295	4,316
Salmonella	1,474	632	2,106
Synechococcus	1,216	521	1,737
Pseudomonas	1,196	513	1,709
Vibrio	1079	463	1,542
Klebsiella	926	397	1,323
Erwinia	667	286	953
Mycobacterium	578	248	826
Staphylococcus	568	244	812
Bacillus	475	204	679

https://doi.org/10.1371/journal.pone.0289030.t003

imposes higher misclassification penalties for minority classes [58, 59]. Formally, let T be the total number of samples in the training set, C be the set of class labels, and t_c be the number of training samples under class $c \in C$. The misclassification penalty w_c for class c is given by Eq (2).

$$w_c = \frac{T}{t_c \cdot |C|} \tag{2}$$

To optimize the weighted F1 score, we conducted hyperparameter tuning with five-fold stratified cross-validation. The hyperparameter space is as follows (the optimal values are in bold): number of trees (50, 100, **150**, 200), number of features to consider in determining the best split ($\log_2$, **square root**), minimum number of samples to split an internal node (**2**, 3, 4), and minimum number of samples to be a leaf node (**1**, 2, 3, 4).

Performance evaluation

In order to factor in our model's confidence in its prediction, we introduced a confidence threshold k. Let p_1 and p_2 be the highest and second-highest predicted class probabilities, respectively, for an input RBP. This input is classified under its predicted class label if and only if $p_1 - p_2 \geq k$. If $p_1 - p_2 < k$, then it is classified as *others* since the model is ambiguous about its classification. S2 Table in S1 File defines the true and false positive and negative outcomes in view of this scheme.

We evaluated our model's performance using weighted precision, recall, F1, and specificity, with the weights corresponding to the class sizes. Formally, let N be the total number of samples in the test set, C be the set of class labels including the *others* class, and n_c be the number of test samples under class $c \in C$. Let $TP_{c,k}$, $TN_{c,k}$, $FP_{c,k}$, and $FN_{c,k}$ be the number of true positive, true negative, false positive, and false negative outcomes for class c at confidence threshold k. The definitions of the evaluation metrics are given in Eqs (3) to (5).

$$\text{Weighted-Precision}_k = \frac{1}{N} \sum_{c \in C} n_c \cdot \frac{TP_{c,k}}{TP_{c,k} + FP_{c,k}} \tag{3}$$

$$\text{Weighted-Recall}_k = \frac{1}{N} \sum_{c \in C} n_c \cdot \frac{TP_{c,k}}{TP_{c,k} + FN_{c,k}} \tag{4}$$

$$\text{Weighted-Specificity}_k = \frac{1}{N} \sum_{c \in C} n_c \cdot \frac{TN_{c,k}}{TN_{c,k} + FP_{c,k}} \tag{5}$$

$$\text{Weighted-F1}_k = \frac{1}{N} \sum_{c \in C} n_c \cdot \frac{2 \cdot \frac{TP_{c,k}}{TP_{c,k} + FP_{c,k}} \cdot \frac{TP_{c,k}}{TP_{c,k} + FN_{c,k}}}{\frac{TP_{c,k}}{TP_{c,k} + FP_{c,k}} + \frac{TP_{c,k}}{TP_{c,k} + FN_{c,k}}} \tag{6}$$

Results

Transformer-based embeddings of receptor-binding proteins outperformed handcrafted genomic and protein features for phage-host interaction prediction

We compared the performance of our model with that of the state-of-the-art phage-host interaction prediction tool by Boeckaerts *et al.* [15]; its feature set consists of 218 handcrafted features, 133 of which are derived from the genomic sequences and 85 from the RBP sequences. To ensure a fair comparison, we retrained this tool on our training dataset. We evaluated the performance across different confidence thresholds ranging from $k = 60\%$ to 100% in steps of 10%.

While representing RBPs via protein embeddings and via these handcrafted features yielded similar weighted precision and specificity scores (S3, S5 Tables in S1 File), the use of protein embeddings improved the weighted F1 and recall across all tested confidence thresholds (Table 4 and S4 Table in S1 File). Among the evaluated protein language models, the highest performance was obtained with the autoencoder model ProtT5, followed by ESM-1b and ESM. In particular, utilizing the ProtT5 embeddings outperformed the handcrafted features by around 3% to 4% in terms of weighted F1 and recall. Meanwhile, the smallest performance increase was registered by the autoregressive model ProtXLNet. The per-class evaluation results are reported in S6 to S10 Tables in S1 File.

Fig 3 plots the weighted precision against the weighted recall scores at confidence thresholds ranging from $k = 0\%$ to 100% in steps of 10%. From $k = 10\%$ to around $k = 50\%$ or 60%, increasing the value of k increases the precision but decreases the recall. Beyond this confidence threshold, increasing the value of k decreases both precision and recall. A reason for this decrease in precision is that setting k to higher (stricter) values results in more samples being mislabeled under the *others* class, which comprises around 12% of our test set. We also provide the weighted precision-recall curve at $k = 0$ (i.e., when the confidence measure is removed and samples are classified according highest class probability resulting in none of the samples, including truly others, as being labeled as *others*) in S1 Fig in S1 File.

Integrating handcrafted features did not significantly increase performance

To determine the extent to which the further integration of handcrafted sequence properties can potentially improve the predictive power of our best-performing model, we combined the ProtT5 embeddings of the RBPs with the sequence properties that registered the highest Gini importance

Table 4. Model performance in terms of weighted F1. The header row refers to the confidence thresholds at which we evaluated model performance; these confidence thresholds range from $k = 60\%$ to 100% in steps of 10%.

	60%	70%	80%	90%	100%
Boeckaerts *et al.* [15]	59.35%	53.41%	47.20%	38.97%	21.61%
SeqVec	60.64%	54.90%	49.12%	40.52%	23.20%
ESM	62.21%	57.13%	51.18%	42.70%	**25.18%**
ESM-1b	62.27%	57.16%	51.29%	42.66%	25.04%
ProtBert	60.58%	55.27%	49.29%	41.07%	24.59%
ProtXLNet	60.18%	54.54%	48.46%	40.20%	23.92%
ProtAlbert	60.45%	54.82%	48.51%	40.57%	23.80%
ProtT5	**62.95%**	**57.51%**	**51.98%**	**43.05%**	24.82%

The highest weighted F1 scores are given in bold and underlined.

https://doi.org/10.1371/journal.pone.0289030.t004

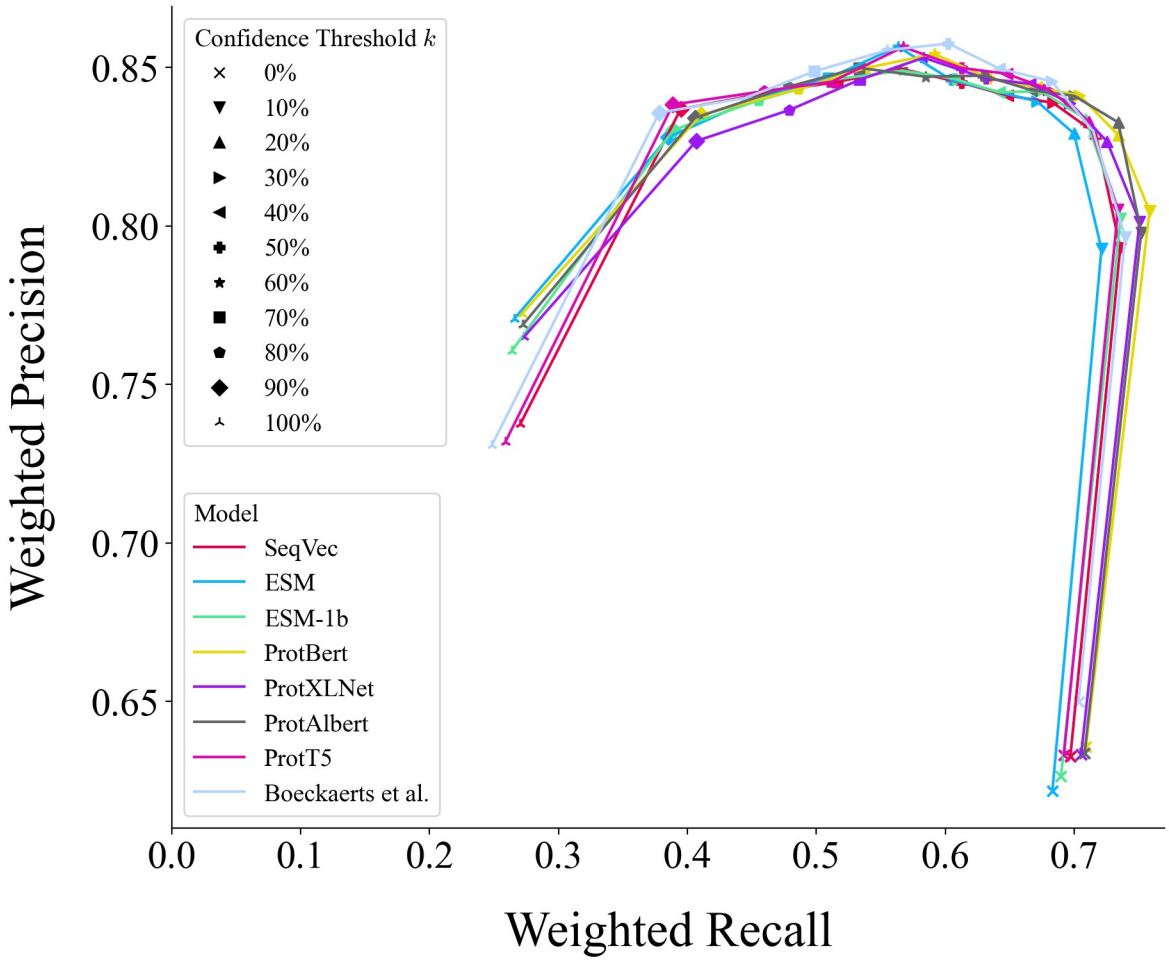

Fig 3. Weighted precision-recall curves showing the model performance. The curves plot the weighted precision against the weighted recall at different confidence thresholds ranging from k = 0% to 100% in steps of 10%.

https://doi.org/10.1371/journal.pone.0289030.g003

after training the model by Boeckaerts *et al.* [15] on our dataset. We also examined the performance if the sequence properties were limited to those extracted from protein sequences.

As reported in Tables 5 to 8, the increase in the weighted F1 scores after integrating these handcrafted sequence properties was consistently below 1.33% across all tested confidence

Table 5. Weighted F1 scores after integrating handcrafted sequence properties to the vector representations of the receptor-binding proteins. The selected sequence properties are those with the highest Gini importance after training the phage-host interaction prediction tool by Boeckaerts *et al.* [15] on our dataset. The header row refers to the confidence thresholds at which we evaluated model performance.

	60%	70%	80%	90%	100%
ProtT5	62.95%	57.51%	51.98%	43.05%	24.82%
ProtT5 + A Nucleotide Frequency	63.49%	58.36%	52.56%	43.37%	24.93%
ProtT5 + GC Content	63.91%	58.60%	52.60%	43.99%	25.57%
ProtT5 + C Nucleotide Frequency	63.57%	58.47%	52.44%	43.48%	25.09%
ProtT5 + TTA Codon Frequency	63.25%	57.60%	51.75%	43.01%	24.97%
ProtT5 + TTA Codon Usage Bias	63.22%	57.56%	51.70%	43.05%	24.63%

https://doi.org/10.1371/journal.pone.0289030.t005

thresholds. Furthermore, the increase in precision, recall, and specificity was also consistently below 1.40% (S11 to S14 Tables in S1 File), 1.41% (S15 to S18 Tables in S1 File), respectively, and 0.18% (S19 to S22 Tables in S1 File). Along with the weighted precision-recall curves in S2 to S5 Figs in S1 File, these results suggest that the ProtT5 embeddings may already be capturing these sequence properties.

Table 6. Weighted F1 scores after integrating handcrafted protein sequence properties to the vector representations of the receptor-binding proteins. The selected protein sequence properties are those with the highest Gini importance after training the phage-host interaction prediction tool by Boeckaerts *et al.* [15] on our dataset. The header row refers to the confidence thresholds at which we evaluated model performance.

	60%	70%	80%	90%	100%
ProtT5	62.95%	57.51%	51.98%	43.05%	24.82%
ProtT5 + K (Lysine) Frequency	63.00%	57.49%	52.00%	43.13%	25.07%
ProtT5 + Isoelectric Point (pI)	62.90%	57.36%	51.76%	43.01%	25.27%
ProtT5 + Fourth Protein Z-Scale*	62.86%	57.40%	51.48%	43.06%	24.97%
ProtT5 + % of Exposed SA†	62.78%	57.65%	51.64%	42.91%	25.42%
ProtT5 + Molecular Weight	62.98%	57.74%	51.79%	42.87%	25.15%

* The fourth protein Z-scale (Z4) [60] is related to the heat of formation, hardness, electronegativity, and electrophilicity.

† *% of Exposed SA* refers to the percentage of residues with exposed solvent accessibility.

https://doi.org/10.1371/journal.pone.0289030.t006

Table 7. Weighted F1 scores after integrating the top *n* handcrafted sequence properties to the vector representations of the receptor-binding proteins. The selected sequence properties are those with the highest Gini importance after training the phage-host interaction prediction tool by Boeckaerts *et al.* [15] on our dataset. These properties (in order of decreasing importance) are the A nucleotide frequency, GC content, C nucleotide frequency, TTA codon frequency, and TTA codon usage bias. The header row refers to the confidence thresholds at which we evaluated model performance.

	60%	70%	80%	90%	100%
ProtT5	62.95%	57.51%	51.98%	43.05%	24.82%
ProtT5 + Top 1	63.49%	58.36%	52.56%	43.37%	24.93%
ProtT5 + Top 2	64.26%	58.76%	52.70%	43.75%	25.13%
ProtT5 + Top 3	64.17%	58.83%	52.81%	44.20%	25.33%
ProtT5 + Top 4	64.03%	58.83%	52.67%	43.85%	24.44%
ProtT5 + Top 5	64.01%	58.68%	52.77%	44.21%	24.80%

https://doi.org/10.1371/journal.pone.0289030.t007

Table 8. Weighted F1 scores after integrating the top *n* handcrafted protein sequence properties to the vector representations of the receptor-binding proteins. The selected protein sequence properties are those with the highest Gini importance after training the phage-host interaction prediction tool by Boeckaerts *et al.* [15] on our dataset. These properties (in order of decreasing importance) are the K (lysine) frequency, isoelectric point, fourth protein Z-scale [60] (which is related to the heat of formation, hardness, electronegativity, and electrophilicity), percentage of residues with exposed solvent accessibility, and molecular weight. The header row refers to the confidence thresholds at which we evaluated model performance.

	60%	70%	80%	90%	100%
ProtT5	62.95%	57.51%	51.98%	43.05%	24.82%
ProtT5 + Protein Top 1	63.00%	57.49%	52.00%	43.13%	25.07%
ProtT5 + Protein Top 2	62.96%	57.57%	51.79%	42.87%	25.17%
ProtT5 + Protein Top 3	62.95%	57.47%	51.38%	43.09%	25.19%
ProtT5 + Protein Top 4	62.92%	57.63%	51.67%	43.36%	25.34%
ProtT5 + Protein Top 5	62.79%	57.82%	51.83%	42.86%	24.80%

https://doi.org/10.1371/journal.pone.0289030.t008

Discussion

Representing receptor-binding proteins using protein language models

The main novelty of our study lies in our use of protein language models to obtain dense vector encodings of receptor-binding proteins, which are known to be major determinants of phage-host specificity [26–29].

Vectorizing sequences by representation learning requires only the sequences themselves, discarding the need to derive additional alignment or structural information and eliminating the difficulty of selecting from a wide array of potentially informative signals [30]. Aside from these general advantages of representation learning over manual feature engineering, our experiments also showed that the use of protein embeddings improves phage-host interaction prediction and outperforms handcrafted genomic and protein features. In particular, utilizing the transformer-based autoencoder model ProtT5 resulted in the best performance, increasing the weighted F1 and recall scores by 3% to 4% across all tested confidence thresholds.

Our embeddings-based model captures a complex combination of features

While protein embeddings have been shown to improve performance in several prototypical bioinformatics tasks [34, 61, 62], their interpretability remains a challenge [63]. The most common approach is to project the embeddings onto a low-dimensional space via nonlinear dimensionality reduction techniques [30, 31, 36, 64], such as *t*-distributed stochastic neighbor embedding (*t*-SNE) [65] and uniform manifold approximation and projection (UMAP) [66], and check for the presence of formed clusters.

These approaches have been employed to establish that protein embeddings carry information on salient physicochemical, structural, and functional properties of protein sequences [33]. For instance, SeqVec and the ProtTrans and ESM families of language models, which we explored in our work, have been shown to capture properties such as hydrophobicity, charge, polarity, and molecular weight [31, 36, 38]. In our experiments, we found no significant performance improvement after combining the ProtT5 embeddings with selected handcrafted features, suggesting that the embeddings may already be capturing guanine-cytosine content [15, 67–69], codon usage bias [15, 70–72], and other important signals of phage-host interaction that emerge from the close coexistence and coevolution of phages and their bacterial hosts.

To visualize the geometry of the embeddings and attempt to identify the biophysical features that they capture in the context of phage-host interaction prediction, we represented each RBP as a vector whose components are the ℓ components of its ProtT5 embedding with the highest Gini importance after training our model. We then employed *t*-SNE and UMAP [66] to project the set of these vectors onto a two-dimensional space. The *t*-SNE projections were generated by setting the perplexity to 50, the number of iterations to 1000, the initialization to principal component analysis, and the learning rate to the number of samples divided by 12 (following Kobak and Berens [73]). The UMAP projections were generated by setting the number of neighbors to 100 and the minimum distance between embedded points to 0.7. We colored the points based on the sequence properties with the highest Gini importance after training the model by Boeckaerts *et al.* [15] on our dataset. Figs 4 and 5 show the resulting plots when ℓ is set to 100.

Experimenting with different values of ℓ, we observed that, although there were instances wherein local clusters appeared to form, the points generally did not show clear separation boundaries, suggesting that the components of the embeddings do not squarely correspond to the individual sequence features under consideration.

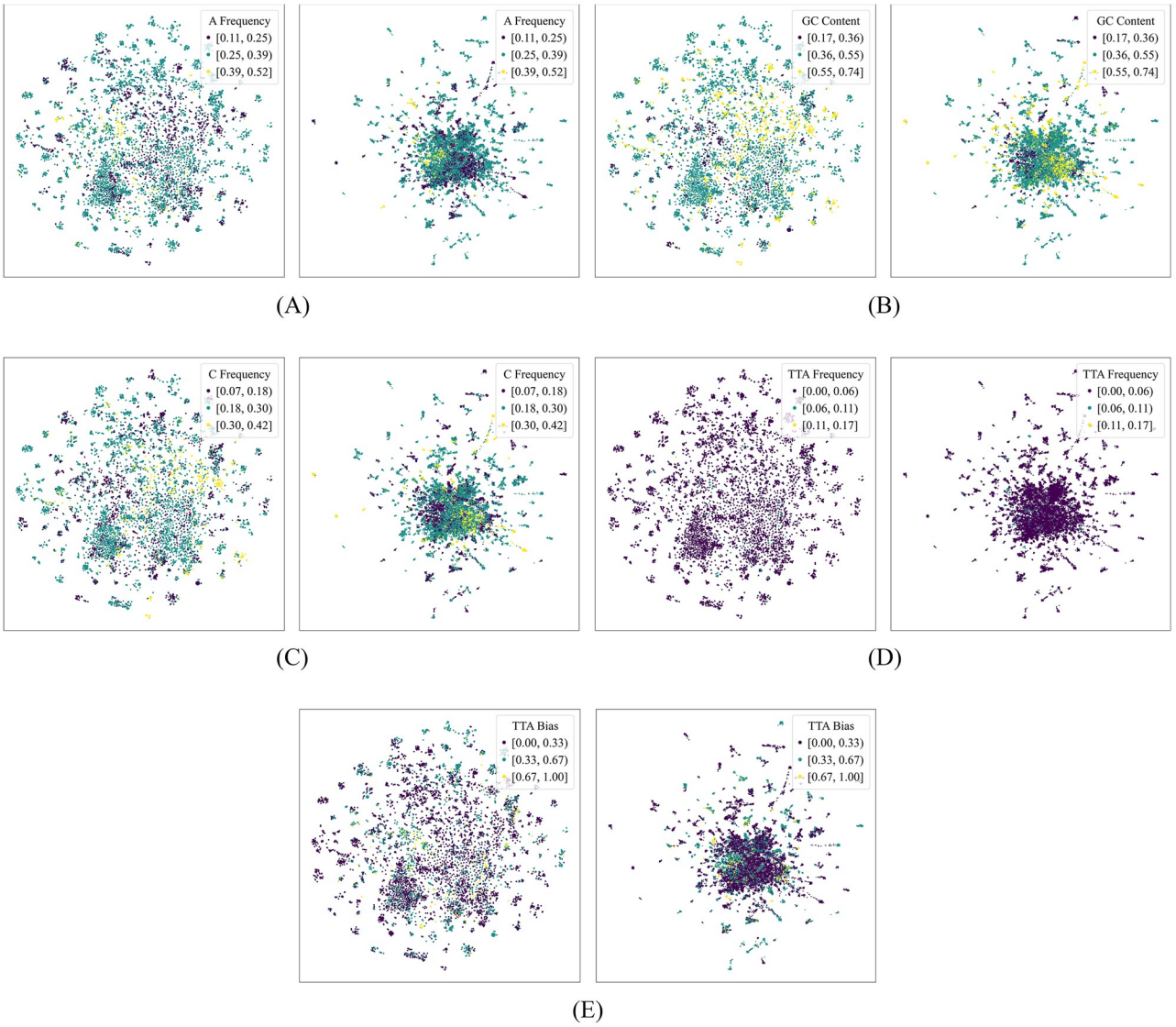

Fig 4. *t*-distributed stochastic neighbor embedding (*t*-SNE) and uniform manifold approximation and projection (UMAP) plots of the ProtT5 embeddings, colored based on handcrafted sequence properties. This figure shows the *t*-SNE (left of each subfigure) and UMAP (right of each subfigure) projections. Each point corresponds to the two-dimensional projection of a subvector of a receptor-binding protein's ProtT5 embedding, the components of which are the ℓ components with the highest Gini importance after training our phage-host interaction prediction model (in this figure, $\ell = 100$). The points were colored based on the sequence properties with the highest Gini importance after training the model by Boeckaerts *et al.* [15] on our dataset; these properties are as follows: (A) A nucleotide frequency, (B) GC content, (C) C nucleotide frequency, (D) TTA codon frequency, and (E) TTA codon usage bias (TTA codes for the amino acid leucine).

https://doi.org/10.1371/journal.pone.0289030.g004

Investigating the exact fashion in which these features are captured in the latent embedding space is a possible research direction, especially as improving the interpretability of deep learning models remains an open challenge. Vig *et al.* [74] viewed the attention weights across all the heads in a layer as a feature vector and fed this vector to classifiers designed for different probing tasks in order to determine whether the biophysical properties being probed are learned in the layer of interest. Another approach employed in previous studies [75, 76] is plotting heatmaps derived from the attention matrix to visualize and examine the attention

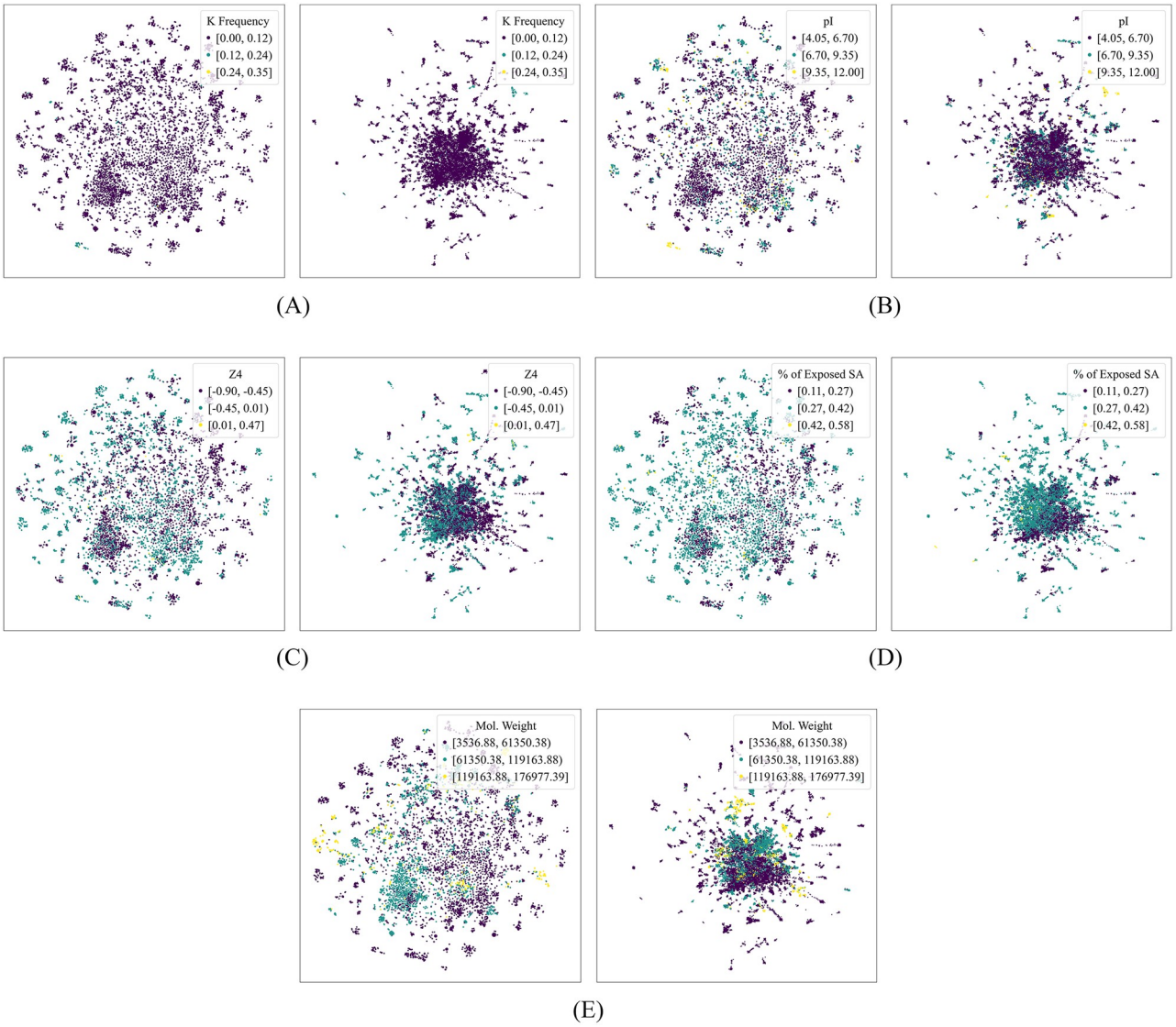

Fig 5. *t*-distributed stochastic neighbor embedding (*t*-SNE) and uniform manifold approximation and projection (UMAP) plots of the ProtT5 embeddings, colored based on handcrafted protein sequence properties. This figure shows the *t*-SNE (left of each subfigure) and UMAP (right of each subfigure) projections. Each point corresponds to the two-dimensional projection of a subvector of a receptor-binding protein's ProtT5 embedding, the components of which are the ℓ components with the highest Gini importance after training our phage-host interaction prediction model (in this figure, $\ell = 100$). The points were colored based on the protein sequence properties with the highest Gini importance after training the model by Boeckaerts *et al.* [15] on our dataset; these properties are as follows: (A) lysine frequency, (B) isoelectric point, (C) fourth protein Z-scale [60] (which is related to the heat of formation, hardness, electronegativity, and electrophilicity), (D) percentage of residues with exposed solvent accessibility, and (E) molecular weight.

https://doi.org/10.1371/journal.pone.0289030.g005

weights. Recently, Hou *et al.* [77] used BertViz [78] to construct bipartite graphs that show the associations among the amino acids in the input sequence in view of self-attention. Utilizing other visualization tools from natural language processing, such as exBERT [79] and AttViz [80], is thus a promising direction that may be explored. Although it is not straightforward to infer the biophysical properties captured in the embeddings based only on these visualizations, they reveal key insights into the internal representations of sequences in transformer models, which, in turn, may serve as bases for biological hypotheses.

Other formulations of the phage-host interaction prediction problem

Our study is primarily interested in investigating how the representation of phages' receptor-binding proteins affects phage-host interaction prediction. To give emphasis on the representation of RBPs, we framed phage-host interaction prediction as a multiclass classification problem with the dense RBP embeddings as the input and the host genus as the output. We also benchmarked our work against a state-of-the-art method [15] that follows the same multiclass formulation but takes a different set of RBP properties (i.e., handcrafted genomic and protein features) as input.

A different formulation of the phage-host interaction prediction problem is as a binary classification problem, where the input is some representation of both phage and host sequences and the output is the presence (or absence) of interaction [14, 16, 20]. It can also be treated as a multiclass classification problem with some representation of the phage sequence as the input and the host as the output [17–19, 22, 23, 81]. Tools that follow this formulation typically have an under-the-hood dataset of information derived from calculating feature scores between phage sequences and a predetermined set of host sequences.

Moreover, while most studies [14, 16, 18–21, 23] consider the entire proteome or genome, our work and that of Boeckaerts *et al.* [15] narrow the input to some representation of selected phage proteins of biological interest (i.e., RBPs). Hence, instead of directly mapping the phages (given their genomes or proteomes) to putative hosts, we perform protein-to-host mapping, i.e., we map the RBPs to putative hosts of the phages from which these proteins were obtained.

While it is not possible to make a direct comparison of our results to the ones using a different approach than ours, we provide in Table 9 a summary of the computational problem formulation and performance of existing machine learning and deep learning tools for phage-host interaction prediction.

Conclusion

In this study, we capitalized on representation learning to automatically encode receptor-binding protein sequences into meaningful dense embeddings. To this end, we extensively tested different protein language models and built a random forest model for phage-host interaction prediction. Our experiments showed that the use of embeddings of receptor-binding proteins presents improvements over handcrafted genomic and protein sequence features, with the highest performance obtained using the transformer-based autoencoder model ProtT5. Moreover, these protein embeddings are able to capture complex combinations of biological features given only the raw sequences, without the need to supply additional alignment or structural information.

Our work makes the simplifying assumption that all the RBPs of a given phage are specific to one host. Albeit significantly less common than single-host phages, some phages are known to possess multiple RBPs, with the RBPs possibly adsorbing with different bacteria. For example, the polyvalent bacteriophage ΦK64–1 has eleven known RBPs targeting a wide spectrum of *Klebsiella* capsular types [91]. However, to the best of our knowledge, there are currently no existing datasets that map individual RBPs to their target hosts.

The receptor-binding proteins considered in our study are also limited to those of tailed phages belonging to the order *Caudovirales*, which constitute around 96% of all known phages [92]. It may also be interesting to explore a similar approach for the computational prediction of interaction between non-tailed phages and their hosts.

Further future directions include improving the interpretability of protein embeddings and incorporating other mechanisms related to phage-host interaction (e.g., restriction-

Table 9. Comparison of existing machine learning and deep learning tools for predicting phage-host interaction.

	Input	Host Level	Test Dataset*	Performance
Binary Classification (Entire Sequence as Input)				
Leite *et al.* [14]	Phage and host proteome	Strain	PhagesDB [82] + GenBank [47]	F1 = 95.9%
PHISDetector† [16]	Phage and host genome	Species	VHM dataset [10]	Accuracy = 51%
PredPHI [20]	Phage and host proteome	Species	PhagesDB [82] + GenBank [47]	AUC-ROC = 81%
PHIAF [21]	Phage and host proteome and genome	Species	PhagesDB [82] + RefSeq [83] + MVP [84] + VHDB [85]	AUC-ROC = 88%
Multiclass Classification (Entire Sequence as Input)				
VirHostMatcher-Net [18]	Phage genome (or contigs)	Genus	RefSeq [83]	Accuracy = 59%
PHP [17]	Phage genome (or contigs)	Genus	VHM dataset [10], NCBI Genome [86]	Accuracy = 34% (VHM dataset), 35% (NCBI Genome)
RaFAH [19]	Phage genome (or contigs)	Genus	RefSeq [83]	F1 = 59%
HoPhage [22]	Phage genome (or contigs)	Genus	RefSeq [83] + VHDB [85]	Accuracy = 81.11%
HostG [81]	Phage genome (or contigs)	Genus	PHP dataset [17]	100% accuracy at softmax threshold of 94%
DeepHost [23]	Phage genome	Species	NCBI Genome [86] + EMBL [87] + PhagesDB [82] + Phage evolution database [88]	Accuracy = 90.78%
Multiclass Classification (Selected Proteins as Input)				
Boeckaerts *et al.* [15]	Phage receptor-binding proteins	Species	UniProtKB [89] + UniRef [40] + Millardlab dataset [90]	AUPR between 73.6% and 93.8% under different sequence similarity thresholds

* Datasets separated by a plus (+) indicate that selected entries from these datasets were aggregated to create a single test set.

† PHISDetector [16] also provides an option to input only the phage genome.

https://doi.org/10.1371/journal.pone.0289030.t009

modification and CRISPR-Cas systems), as well as host sequence information possibly encoded as dense embeddings.

Glossary

- Autoencoding transformer—A type of transformer that is pretrained in a "bidirectional" fashion, i.e., it can read all the input tokens during pretraining
- Autoregressive transformer—A type of transformer that is pretrained in a "unidirectional" fashion, i.e., it attempts to predict the next token given only the previous tokens
- Extreme gradient boosting—A machine learning algorithm that iteratively combines weak "learners" (in this case, decision trees) to improve performance
- Hidden Markov model—A statistical model for modeling systems (including biological sequences) characterized by a sequence of observations and unobserved (hidden) states
- Protein language model—A deep learning model that adopts the architecture of models from natural language processing in order to convert protein sequences into dense vector representations (embeddings)
- Random forest—A machine learning algorithm that employs an ensemble of multiple decision trees to produce an output

- RBP—Receptor-binding protein. A specific type of protein found in tailed phages that is responsible for initiating the recognition and infection of bacterial hosts
- Transformer—A deep learning architecture characterized by the use of attention mechanisms to improve performance
- *t*-SNE—*t*-distributed stochastic neighbor embedding. A nonlinear method for projecting high-dimensional vectors onto a low-dimensional space by calculating a joint probability distribution that captures the similarity between data points
- UMAP—Uniform manifold approximation and projection. A nonlinear method for projecting high-dimensional vectors onto a low-dimensional space by assuming that the data points are evenly distributed on some topological space

Supporting information

S1 File.
(PDF)

Acknowledgments

The authors thank Dr. Paul K. Yu of the Systems and Computational Biology Research Unit, De La Salle University, for his feedback on improving the study's reproducibility.

Author Contributions

Conceptualization: Anish M. S. Shrestha.

Formal analysis: Mark Edward M. Gonzales, Jennifer C. Ureta.

Methodology: Mark Edward M. Gonzales, Anish M. S. Shrestha.

Software: Mark Edward M. Gonzales.

Supervision: Anish M. S. Shrestha.

Validation: Mark Edward M. Gonzales, Jennifer C. Ureta.

Writing – original draft: Mark Edward M. Gonzales.

Writing – review & editing: Mark Edward M. Gonzales, Jennifer C. Ureta, Anish M. S. Shrestha.

References

1. Murray CJ, Ikuta KS, Sharara F, Swetschinski L, Aguilar GR, Gray A, et al. Global burden of bacterial antimicrobial resistance in 2019: A systematic analysis. The Lancet. 2022; 399(10325):629–655. https://doi.org/10.1016/S0140-6736(21)02724-0
2. Robredo JP, Eala M, Paguio JA, Salamat M, Celi L. The challenges of combatting antimicrobial resistance in the Philippines. The Lancet Microbe. 2022; 3:E246. https://doi.org/10.1016/S2666-5247(22)00029-5 PMID: 35544059
3. Pokharel S, Raut S, Adhikari B. Tackling antimicrobial resistance in low-income and middle-income countries. BMJ Global Health. 2019; 4(6). https://doi.org/10.1136/bmjgh-2019-002104 PMID: 31799007
4. Taylor P, Reeder R. Antibiotic use on crops in low and middle-income countries based on recommendations made by agricultural advisors. CABI Agriculture and Bioscience. 2020; 1(1):1. https://doi.org/10.1186/s43170-020-00001-y

5. Mann A, Nehra K, Rana JS, Dahiya T. Antibiotic resistance in agriculture: Perspectives on upcoming strategies to overcome upsurge in resistance. Current Research in Microbial Sciences. 2021; 2:100030. https://doi.org/10.1016/j.crmicr.2021.100030 PMID: 34841321

6. Zhao J, Zhang Z, Tian C, Chen X, Hu L, Wei X, et al. Characterizing the Biology of Lytic Bacteriophage vB_EaeM_φEap-3 Infecting Multidrug-Resistant Enterobacter aerogenes. Front Microbiol. 2019; 10:420. https://doi.org/10.3389/fmicb.2019.00420 PMID: 30891025

7. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. Journal of Molecular Biology. 1990; 215(3):403–410. https://doi.org/10.1016/S0022-2836(05)80360-2 PMID: 2231712

8. Zielezinski A, Barylski J, Karlowski WM. Taxonomy-aware, sequence similarity ranking reliably predicts phage–host relationships. BMC Biology. 2021; 19(1):223. https://doi.org/10.1186/s12915-021-01146-6 PMID: 34625070

9. Villarroel J, Kleinheinz KA, Jurtz VI, Zschach H, Lund O, Nielsen M, et al. HostPhinder: A Phage Host Prediction Tool. Viruses. 2016; 8(5):116. https://doi.org/10.3390/v8050116 PMID: 27153081

10. Ahlgren NA, Ren J, Lu YY, Fuhrman JA, Sun F. Alignment-free d_2^* oligonucleotide frequency dissimilarity measure improves prediction of hosts from metagenomically-derived viral sequences. Nucleic Acids Research. 2016; 45(1):39–53. https://doi.org/10.1093/nar/gkw1002 PMID: 27899557

11. Galiez C, Siebert M, Enault F, Vincent J, Söding J. WIsH: Who is the host? Predicting prokaryotic hosts from metagenomic phage contigs. Bioinformatics. 2017; 33(19):3113–3114. https://doi.org/10.1093/bioinformatics/btx383 PMID: 28957499

12. Ofir G, Sorek R. Contemporary Phage Biology: From Classic Models to New Insights. Cell. 2018; 172 (6):1260–1270. https://doi.org/10.1016/j.cell.2017.10.045 PMID: 29522746

13. Versoza CJ, Pfeifer SP. Computational Prediction of Bacteriophage Host Ranges. Microorganisms. 2022; 10(1). https://doi.org/10.3390/microorganisms10010149 PMID: 35056598

14. Leite D, Lopez J, Brochet X, Barreto-Sanz M, Que Y, Resch G, et al. Exploration of multiclass and one-class learning methods for prediction of phage-bacteria interaction at strain level. In: 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). Los Alamitos, CA, USA: IEEE Computer Society; 2018. p. 1818–1825.

15. Boeckaerts D, Stock M, Criel B, Gerstmans H, De Baets B, Briers Y. Predicting bacteriophage hosts based on sequences of annotated receptor-binding proteins. Scientific Reports. 2021; 11(1):1467. https://doi.org/10.1038/s41598-021-81063-4 PMID: 33446856

16. Zhou F, Gan R, Zhang F, Ren C, Yu L, Si Y, et al. PHISDetector: A tool to detect diverse in silico phage–host interaction signals for virome studies. Genomics, Proteomics & Bioinformatics. 2022. https://doi.org/10.1016/j.gpb.2022.02.003 PMID: 35272051

17. Lu C, Zhang Z, Cai Z, Zhu Z, Qiu Y, Wu A, et al. Prokaryotic virus host predictor: A Gaussian model for host prediction of prokaryotic viruses in metagenomics. BMC Biology. 2021; 19(1):5. https://doi.org/10.1186/s12915-020-00938-6 PMID: 33441133

18. Wang W, Ren J, Tang K, Dart E, Ignacio-Espinoza JC, Fuhrman J, et al. A network-based integrated framework for predicting virus–prokaryote interactions. NAR Genomics and Bioinformatics. 2020; 2(2). https://doi.org/10.1093/nargab/lqaa044 PMID: 32626849

19. Coutinho FH, Zaragoza-Solas A, López-Pérez M, Barylski J, Zielezinski A, Dutilh BE, et al. RaFAH: Host prediction for viruses of Bacteria and Archaea based on protein content. Patterns (N Y). 2021; 2 (7):100274. https://doi.org/10.1016/j.patter.2021.100274 PMID: 34286299

20. Li M, Wang Y, Li F, Zhao Y, Liu M, Zhang S, et al. A Deep Learning-Based Method for Identification of Bacteriophage-Host Interaction. IEEE/ACM Transactions on Computational Biology and Bioinformatics. 2021; 18(5):1801–1810. https://doi.org/10.1109/TCBB.2020.3017386 PMID: 32813660

21. Li M, Zhang W. PHIAF: Prediction of phage-host interactions with GAN-based data augmentation and sequence-based feature fusion. Briefings in Bioinformatics. 2021; 23(1).

22. Tan J, Fang Z, Wu S, Guo Q, Jiang X, Zhu H. HoPhage: An ab initio tool for identifying hosts of phage fragments from metaviromes. Bioinformatics. 2021; 38(2):543–545. https://doi.org/10.1093/bioinformatics/btab585

23. Ruohan W, Xianglilan Z, Jianping W, Shuai Cheng LI. DeepHost: Phage host prediction with convolutional neural network. Briefings in Bioinformatics. 2021; 23(1). https://doi.org/10.1093/bib/bbab385

24. Häuser R, Blasche S, Dokland T, Haggård-Ljungquist E, von Brunn A, Salas M, et al. Bacteriophage protein-protein interactions. Adv Virus Res. 2012; 83:219–298. https://doi.org/10.1016/B978-0-12-394438-2.00006-2 PMID: 22748812

25. Nobrega FL, Vlot M, de Jonge PA, Dreesens LL, Beaumont HJE, Lavigne R, et al. Targeting mechanisms of tailed bacteriophages. Nature Reviews Microbiology. 2018; 16(12):760–773. https://doi.org/10.1038/s41579-018-0070-8 PMID: 30104690

26. Guerrero-Ferreira RC, Viollier PH, Ely B, Poindexter JS, Georgieva M, Jensen GJ, et al. Alternative mechanism for bacteriophage adsorption to the motile bacterium Caulobacter crescentus. Proceedings of the National Academy of Sciences. 2011; 108(24):9963–9968. https://doi.org/10.1073/pnas.1012388108

27. Zampara A, Sørensen MCH, Grimon D, Antenucci F, Vitt AR, Bortolaia V, et al. Exploiting phage receptor binding proteins to enable endolysins to kill Gram-negative bacteria. Scientific Reports. 2020; 10 (1):12087. https://doi.org/10.1038/s41598-020-68983-3 PMID: 32694655

28. Santos SB, Cunha AP, Macedo M, Nogueira CL, Brandão A, Costa SP, et al. Bacteriophage-receptor binding proteins for multiplex detection of Staphylococcus and Enterococcus in blood. Biotechnology and Bioengineering. 2020; 117(11):3286–3298. https://doi.org/10.1002/bit.27489 PMID: 32658303

29. Tremblay DM, Tegoni M, Spinelli S, Campanacci V, Blangy S, Huyghe C, et al. Receptor-Binding Protein of Lactococcus lactis Phages: Identification and Characterization of the Saccharide Receptor-Binding Site. Journal of Bacteriology. 2006; 188(7):2400–2410. https://doi.org/10.1128/JB.188.7.2400-2410.2006 PMID: 16547026

30. Yang KK, Wu Z, Bedbrook CN, Arnold FH. Learned protein embeddings for machine learning. Bioinformatics. 2018; 34(15):2642–2648. https://doi.org/10.1093/bioinformatics/bty178 PMID: 29584811

31. Elnaggar A, Heinzinger M, Dallago C, Rehawi G, Wang Y, Jones L, et al. ProtTrans: Toward Understanding the Language of Life Through Self-Supervised Learning. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2022; 44(10):7112–7127. https://doi.org/10.1109/TPAMI.2021.3095381 PMID: 34232869

32. Littmann M, Heinzinger M, Dallago C, Weissenow K, Rost B. Protein embeddings and deep learning predict binding residues for various ligand classes. Scientific Reports. 2021; 11(1):23916. https://doi.org/10.1038/s41598-021-03431-4 PMID: 34903827

33. Bepler T, Berger B. Learning the protein language: Evolution, structure, and function. Cell Systems. 2021; 12(6):654–669.e3. https://doi.org/10.1016/j.cels.2021.05.017 PMID: 34139171

34. Pokharel S, Pratyush P, Heinzinger M, Newman RH, KC DB. Improving protein succinylation sites prediction using embeddings from protein language model. Scientific Reports. 2022; 12(1):16933. https://doi.org/10.1038/s41598-022-21366-2 PMID: 36209286

35. Marquet C, Heinzinger M, Olenyi T, Dallago C, Erckert K, Bernhofer M, et al. Embeddings from protein language models predict conservation and variant effects. Human Genetics. 2022; 141(10):1629–1647. https://doi.org/10.1007/s00439-021-02411-y PMID: 34967936

36. Heinzinger M, Elnaggar A, Wang Y, Dallago C, Nechaev D, Matthes F, et al. Modeling aspects of the language of life through transfer-learning protein sequences. BMC Bioinformatics. 2019; 20(1):723. https://doi.org/10.1186/s12859-019-3220-8 PMID: 31847804

37. Peters ME, Neumann M, Iyyer M, Gardner M, Clark C, Lee K, et al. Deep Contextualized Word Representations. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). New Orleans, Louisiana: Association for Computational Linguistics; 2018. p. 2227–2237.

38. Rives A, Meier J, Sercu T, Goyal S, Lin Z, Liu J, et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. Proceedings of the National Academy of Sciences. 2021; 118(15):e2016239118. https://doi.org/10.1073/pnas.2016239118 PMID: 33876751

39. Apweiler R, Bairoch A, Wu C, Barker W, Boeckmann B, Ferro S, et al. UniProt: The Universal Protein Knowledgebase. Nucleic acids research. 2004; 32:D115–9. https://doi.org/10.1093/nar/gkh131 PMID: 14681372

40. Suzek BE, Huang H, McGarvey P, Mazumder R, Wu CH. UniRef: Comprehensive and non-redundant UniProt reference clusters. Bioinformatics. 2007; 23(10):1282–1288. https://doi.org/10.1093/bioinformatics/btm098 PMID: 17379688

41. Steinegger M, Mirdita M, Söding J. Protein-level assembly increases protein sequence recovery from metagenomic samples manyfold. Nat Methods. 2019; 16(7):603–606. https://doi.org/10.1038/s41592-019-0437-4 PMID: 31235882

42. Steinegger M, Söding J. Clustering huge protein sequence sets in linear time. Nature Communications. 2018; 9(1):2542. https://doi.org/10.1038/s41467-018-04964-5 PMID: 29959318

43. Freepik. Bacteria free icon;. https://cdn-icons-png.flaticon.com/512/112/112736.png.

44. Freepik. Protein free icon;. https://cdn-icons-png.flaticon.com/512/1951/1951420.png.

45. Becris. Deep learning free icon;. https://cdn-icons-png.flaticon.com/512/2103/2103718.png.

46. Cook R, Brown N, Redgwell T, Rihtman B, Barnes M, Clokie M, et al. INfrastructure for a PHAge REference Database: Identification of Large-Scale Biases in the Current Collection of Cultured Phage Genomes. PHAGE. 2021; 2(4):214–223. https://doi.org/10.1089/phage.2021.0007 PMID: 36159887

47. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL. GenBank. Nucleic Acids Research. 2006; 35(suppl_1):D21–D25. https://doi.org/10.1093/nar/gkj157 PMID: 16381837

48. Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics. 2009; 25(11):1422–1423. https://doi.org/10.1093/bioinformatics/btp163 PMID: 19304878

49. Schoch CL, Ciufo S, Domrachev M, Hotton CL, Kannan S, Khovanskaya R, et al. NCBI Taxonomy: a comprehensive update on curation, resources and tools. Database. 2020; 2020. https://doi.org/10.1093/database/baaa062 PMID: 32761142

50. Boeckaerts D, Stock M, De Baets B, Briers Y. Identification of Phage Receptor-Binding Protein Sequences with Hidden Markov Models and an Extreme Gradient Boosting Classifier. Viruses. 2022; 14(6). https://doi.org/10.3390/v14061329

51. Seemann T. Prokka: rapid prokaryotic genome annotation. Bioinformatics. 2014; 30(14):2068–2069. https://doi.org/10.1093/bioinformatics/btu153 PMID: 24642063

52. Hyatt D, Chen GL, LoCascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics. 2010; 11(1):119. https://doi.org/10.1186/1471-2105-11-119 PMID: 20211023

53. Terzian P, Olo Ndela E, Galiez C, Lossouarn J, Pérez Bucio R, Mom R, et al. PHROG: families of prokaryotic virus proteins clustered using remote homology. NAR Genomics and Bioinformatics. 2021; 3 (3). https://doi.org/10.1093/nargab/lqab067 PMID: 34377978

54. Billaud M, Lamy-Besnier Q, Lossouarn J, Moncaut E, Dion MB, Moineau S, et al. Analysis of viromes and microbiomes from pig fecal samples reveals that phages and prophages rarely carry antibiotic resistance genes. ISME Communications. 2021; 1(1):55. https://doi.org/10.1038/s43705-021-00054-8

55. Muscatt G, Hilton S, Raguideau S, Teakle G, Lidbury IDEA, Wellington EMH, et al. Crop management shapes the diversity and activity of DNA and RNA viruses in the rhizosphere. Microbiome. 2022; 10 (1):181. https://doi.org/10.1186/s40168-022-01371-3 PMID: 36280853

56. Dallago C, Schütze K, Heinzinger M, Olenyi T, Littmann M, Lu AX, et al. Learned Embeddings from Deep Learning to Visualize and Predict Protein Sets. Current Protocols. 2021; 1(5):e113. https://doi.org/10.1002/cpz1.113 PMID: 33961736

57. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research. 2011; 12:2825–2830.

58. Khalilia M, Chakraborty S, Popescu M. Predicting disease risks from highly imbalanced data using random forest. BMC Medical Informatics and Decision Making. 2011; 11(1):51. https://doi.org/10.1186/1472-6947-11-51 PMID: 21801360

59. Chen C, Liaw A, Breiman L. Using random forest to learn imbalanced data. Dept. Statistics. Univ California, Berkeley, CA, Tech Rep. 2004;666.

60. Sandberg M, Eriksson L, Jonsson J, Sjöström M, Wold S. New chemical descriptors relevant for the design of biologically active peptides. A multivariate characterization of 87 amino acids. J Med Chem. 1998; 41(14):2481–2491. https://doi.org/10.1021/jm9700575 PMID: 9651153

61. Iuchi H, Matsutani T, Yamada K, Iwano N, Sumi S, Hosoda S, et al. Representation learning applications in biological sequence analysis. Comput Struct Biotechnol J. 2021; 19:3198–3208. https://doi.org/10.1016/j.csbj.2021.05.039 PMID: 34141139

62. Villegas-Morcillo A, Makrodimitris S, van Ham RCHJ, Gomez AM, Sanchez V, Reinders MJT. Unsupervised protein embeddings outperform hand-crafted sequence and structure features at predicting molecular function. Bioinformatics. 2020; 37(2):162–170. https://doi.org/10.1093/bioinformatics/btaa701

63. Ofer D, Brandes N, Linial M. The language of proteins: NLP, machine learning & protein sequences. Computational and Structural Biotechnology Journal. 2021; 19:1750–1758. https://doi.org/10.1016/j.csbj.2021.03.022 PMID: 33897979

64. Odrzywolek K, Karwowska Z, Majta J, Byrski A, Milanowska-Zabel K, Kosciolek T. Deep embeddings to comprehend and visualize microbiome protein space. Scientific Reports. 2022; 12(1):10332. https://doi.org/10.1038/s41598-022-14055-7 PMID: 35725732

65. van der Maaten L, Hinton G. Visualizing Data using t-SNE. Journal of Machine Learning Research. 2008; 9(86):2579–2605.

66. McInnes L, Healy J, Saul N, Großberger L. UMAP: Uniform Manifold Approximation and Projection. Journal of Open Source Software. 2018; 3(29):861. https://doi.org/10.21105/joss.00861

67. Almpanis A, Swain M, Gatherer D, McEwan N. Correlation between bacterial G+C content, genome size and the G+C content of associated plasmids and bacteriophages. Microb Genom. 2018; 4(4). https://doi.org/10.1099/mgen.0.000168 PMID: 29633935

68. Motlagh AM, Bhattacharjee AS, Coutinho FH, Dutilh BE, Casjens SR, Goel RK. Insights of Phage-Host Interaction in Hypersaline Ecosystem through Metagenomics Analyses. Front Microbiol. 2017; 8:352. https://doi.org/10.3389/fmicb.2017.00352 PMID: 28316597

69. Kortright KE, Chan BK, Turner PE. High-throughput discovery of phage receptors using transposon insertion sequencing of bacteria. Proceedings of the National Academy of Sciences. 2020; 117 (31):18670–18679. https://doi.org/10.1073/pnas.2001888117 PMID: 32675236

70. Lucks JB, Nelson DR, Kudla GR, Plotkin JB. Genome landscapes and bacteriophage codon usage. PLoS Comput Biol. 2008; 4(2):e1000001. https://doi.org/10.1371/journal.pcbi.1000001 PMID: 18463708

71. Ge Z, Li X, Cao X, Wang R, Hu W, Gen L, et al. Viral adaption of staphylococcal phage: A genome-based analysis of the selective preference based on codon usage Bias. Genomics. 2020; 112(6):4657–4665. https://doi.org/10.1016/j.ygeno.2020.08.012 PMID: 32818632

72. Carbone A. Codon bias is a major factor explaining phage evolution in translationally biased hosts. J Mol Evol. 2008; 66(3):210–223. https://doi.org/10.1007/s00239-008-9068-6 PMID: 18286220

73. Kobak D, Berens P. The art of using t-SNE for single-cell transcriptomics. Nature Communications. 2019; 10(1):5416. https://doi.org/10.1038/s41467-019-13056-x PMID: 31780648

74. Vig J, Madani A, Varshney LR, Xiong C, Socher R, Rajani N. BERTology Meets Biology: Interpreting Attention in Protein Language Models. In: International Conference on Learning Representations; 2021.

75. Wang R, Jiang Y, Jin J, Yin C, Yu H, Wang F, et al. DeepBIO: an automated and interpretable deep-learning platform for high-throughput biological sequence prediction, functional annotation and visualization analysis. Nucleic Acids Research. 2023; 51(7):3017–3029. https://doi.org/10.1093/nar/gkad055 PMID: 36796796

76. Yamaguchi H, Saito Y. Evotuning protocols for Transformer-based variant effect prediction on multi-domain proteins. Brief Bioinform. 2021; 22(6). https://doi.org/10.1093/bib/bbab287

77. Hou Z, Yang Y, Ma Z, Wong Kc, Li X. Learning the protein language of proteome-wide protein-protein binding sites via explainable ensemble deep learning. Communications Biology. 2023; 6(1):73. https://doi.org/10.1038/s42003-023-04462-5 PMID: 36653447

78. Vig J. BertViz: A tool for visualizing multihead self-attention in the BERT model. In: ICLR Workshop: Debugging Machine Learning Models; 2019.

79. Hoover B, Strobelt H, Gehrmann S. exBERT: A Visual Analysis Tool to Explore Learned Representations in Transformer Models. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations. Online: Association for Computational Linguistics; 2020. p. 187–196. Available from: https://aclanthology.org/2020.acl-demos.22.

80. Škrlj B, Sheehan S, Eržen N, Robnik-Šikonja M, Luz S, Pollak S. Exploring Neural Language Models via Analysis of Local and Global Self-Attention Spaces. In: Proceedings of the EACL Hackashop on News Media Content Analysis and Automated Report Generation. Online: Association for Computational Linguistics; 2021. p. 76–83. Available from: https://aclanthology.org/2021.hackashop-1.11.

81. Shang J, Sun Y. Predicting the hosts of prokaryotic viruses using GCN-based semi-supervised learning. BMC Biology. 2021; 19(1):250. https://doi.org/10.1186/s12915-021-01180-4 PMID: 34819064

82. Russell DA, Hatfull GF. PhagesDB: The actinobacteriophage database. Bioinformatics. 2016; 33 (5):784–786. https://doi.org/10.1093/bioinformatics/btw711

83. O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, et al. Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. Nucleic Acids Research. 2015; 44(D1):D733–D745. https://doi.org/10.1093/nar/gkv1189 PMID: 26553804

84. Gao NL, Zhang C, Zhang Z, Hu S, Lercher MJ, Zhao XM, et al. MVP: A microbe-phage interaction database. Nucleic Acids Res. 2018; 46(D1):D700–D707. https://doi.org/10.1093/nar/gkx1124 PMID: 29177508

85. Mihara T, Nishimura Y, Shimizu Y, Nishiyama H, Yoshikawa G, Uehara H, et al. Linking Virus Genomes with Host Taxonomy. Viruses. 2016; 8(3):66. https://doi.org/10.3390/v8030066 PMID: 26938550

86. Sayers EW, Agarwala R, Bolton EE, Brister JR, Canese K, Clark K, et al. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. 2019; 47(D1):D23–D28. https://doi.org/10.1093/nar/gky1069 PMID: 30395293

87. Stoesser G, Baker W, van den Broek A, Camon E, Garcia-Pastor M, Kanz C, et al. The EMBL Nucleotide Sequence Database. Nucleic Acids Res. 2002; 30(1):21–26. https://doi.org/10.1093/nar/30.1.21 PMID: 11752244

88. Mavrich TN, Hatfull GF. Bacteriophage evolution differs by host, lifestyle and genome. Nat Microbiol. 2017; 2:17112. https://doi.org/10.1038/nmicrobiol.2017.112 PMID: 28692019

89. The UniProt Consortium. UniProt: The universal protein knowledgebase in 2021. Nucleic Acids Research. 2020; 49(D1):D480–D489. https://doi.org/10.1093/nar/gkaa1100

90. Millardlab. Bacteriophage genomes—April 2019; 2019. https://millardlab.org/home/bacteriophage-genomes/.

91. Pan YJ, Lin TL, Chen CC, Tsai YT, Cheng YH, Chen YY, et al. Klebsiella Phage ΦK64-1 Encodes Multiple Depolymerases for Multiple Host Capsular Types. Journal of virology. 2017; 91(6):e02457–16. https://doi.org/10.1128/JVI.02457-16 PMID: 28077636

92. Taslem Mourosi J, Awe A, Guo W, Batra H, Ganesh H, Wu X, et al. Understanding Bacteriophage Tail Fiber Interaction with Host Surface Receptor: The Key "Blueprint" for Reprogramming Phage Host Range. Int J Mol Sci. 2022; 23(20). https://doi.org/10.3390/ijms232012146 PMID: 36292999