Prediction of virus-host association using protein language models and multiple instance learning

Dan Liu[1], Francesca Young[1,2], David L Robertson[1,#], Ke Yuan[2,3,4,#]

1 MRC−University of Glasgow Center for Virus Research, Glasgow, United Kingdom
2 School of Computing Science, University of Glasgow, Glasgow, United Kingdom
3 School of Cancer Sciences, University of Glasgow, Glasgow, United Kingdom
4 Cancer Research UK Beatson Institute, Glasgow, United Kingdom
#Correspondence: david.l.robertson@glasgow.ac.uk, ke.yuan@glasgow.ac.uk

Abstract

Predicting virus-host association is essential to understand how viruses interact with host species, and discovering new therapeutics for viral diseases across humans and animals. Currently, the host of the majority of viruses is unknown. Here, we introduce EvoMIL, a deep learning method that predicts virus-host association at the species level from viral sequence only. The method combines a pre-trained large protein language model and attention-based multiple instance learning to allow protein-orientated predictions. Our results show that protein embeddings capture stronger predictive signals than traditional handcrafted features, including amino acids and DNA k-mers, and physio-chemical properties. EvoMIL binary classifiers achieve AUC values of over 0.95 for all prokaryotic and nearly 0.8 for almost all eukaryotic hosts. In multi-host prediction tasks, EvoMIL achieved median performance improvements of 8.6% in prokaryotic hosts and 1.8% in eukaryotic hosts. Furthermore, EvoMIL estimates the importance of single proteins in the prediction and maps them to an embedding landscape of all viral proteins, where proteins with similar functions are distinctly clustered together.

Author summary

Being able to predict which viruses can infect which hosts, and identifying the specific proteins that are involved in these interactions, is crucial for understanding viral diseases and developing more effective treatments. Traditional methods for predicting these interactions rely on handcrafted common features among proteins, overlooking the importance of single proteins. We have developed a new method that combines a protein language model and multiple instance learning to allow host prediction directly from protein sequences, without the need to extract handcrafted features. This method significantly improved multiple host association accuracy and revealed the key proteins involved in virus-host interactions.

1 Introduction

Being able to predict which viruses can infect which hosts are a crucial step for understanding how viruses interact with hosts and cause diseases. Despite advances in

sequencing technology, more than 90% of virus sequences held in publicly available databases are not annotated with any host information [1]. Currently, there are no high-throughput experimental methods that can definitively assign a host to these uncultivated viruses. This has in turn limited the options for the treatment and prevention of viral disease, which has been demonstrated by the recent COVID-19 pandemic.

To address this gap, a number of computational approaches have been developed to predict unknown virus-host associations. The coevolution of a virus and its host left signals in their genomes. These signals have been exploited for in silico prediction of virus-host associations in numerous approaches They fall into two broad types: 1) Alignment-based approaches that search for homology such as prophage [2] or CRISPR-cas spacers [3, 4]; 2) Alignment-free methods use features such as k-mer composition, codon usage, or GC content to measure the similarity between viral and host sequences or to other viruses with a known host. To date, there are no computational approaches that consider the structure of proteins from viruses for prediction purposes. Transformer-based protein language models generate protein embeddings that encode structural features based on large-scale protein databases.

Here, we present a virus-host prediction model combining protein language models and multiple instance learning. Transformers are self-supervised deep learning models [5] that learn the semantics of words within a sentence and now dominate the field of Natural Language Processing. More recently, the same architecture has been applied in biology, where words are replaced by amino acids and sentences by protein sequences. Protein language models are trained on publicly available protein sequence archives and learn multi-level biological information from physio-chemical properties of the individual amino acids to structural and functional information about proteins. Multiple instance learning (MIL) is a form of weakly supervised learning that was developed for image processing tasks [6]. Instead of using individually labelled instances for classification, multiple instances are arranged together in a bag with a single label and classified together. We use attention-based MIL, [7] which has the additional advantage of weighting the instances in a bag, thereby indicating the importance of each instance in prediction.

The combination of the two approaches is particularly suited for virus-host prediction where a virus is a bag of different proteins. The proteins collectively contribute to the association with a host. Instead of relying on predefined features, protein language models provide automatically learned features, free from design biases and limitations of the traditional approaches. The enhanced ability to measure similarity and differences between protein sequences further boosts prediction performance through multiple instance learning, where viral proteins enabling interaction with hosts are highlighted through unbiased weight estimation.

In this paper, we introduce EvoMIL a method for predicting virus-host associations by combining the (Evo)lutionary Scale Modeling with (M)multiple (I)instance (L)earning, Fig 1. EvoMIL, uses the model ESM-1b [8] to transform viral protein sequences into embeddings (i.e. numerical vectors) that are then used as features for virus-host classification. Multiple instance learning allows us to treat each virus as a "bag" of proteins. We demonstrate that the embeddings capture the host signal from the viral sequences achieving high prediction scores at the species level of both prokaryotic and eukaryotic hosts. Furthermore, attention-based MIL enables us to identify which proteins are most important in driving prediction and by implication important to virus-host specificity.

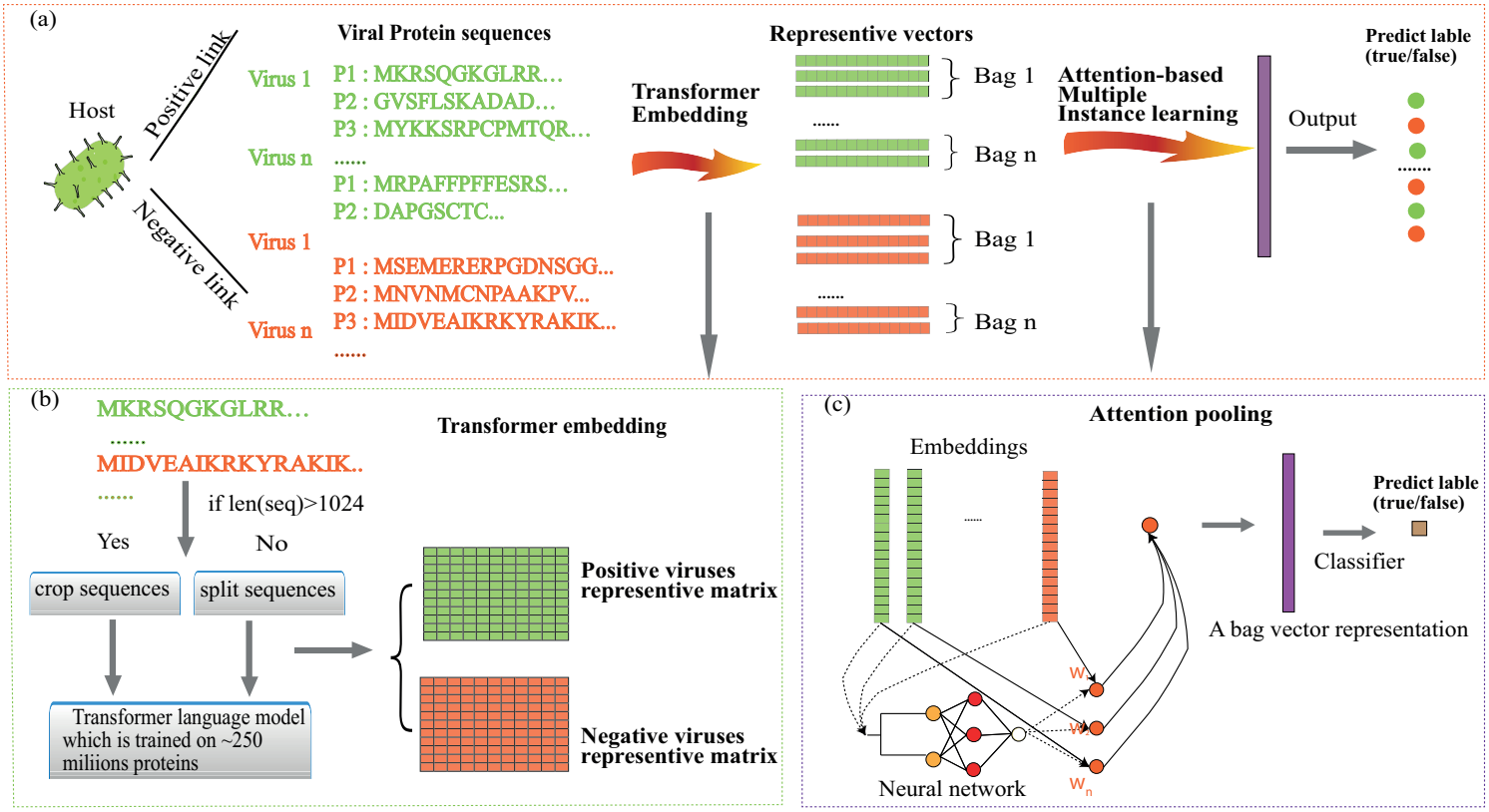

Fig 1. The flow chart of EvoMIL. (a) the whole processes of our approach: firstly, protein sequences of viruses and virus-host links can be collected from the VHDB [9]. For each host, we collect the same number of positive and negative viruses, and then embeddings of protein sequences from viruses are obtained by the pre-trained transformer model [8], which are features for host predictions based on attention-based MIL; (b) protein sequences of viruses are split to sub-sequences for generating embeddings; (c) there is a host label for a set of protein sequences for each virus, attention-based MIL is applied to train the model for each host dataset by protein feature matrices of viruses, finally predict the host label for each virus, and obtain the instance weight for each protein of the virus.

2 Results

2.1 A benchmarking dataset for predicting virus-host association

Balanced binary datasets were generated from known virus-host associations documented in the Virus-Host database, (VHDB), [9] for all hosts with a minimum threshold number of associations. These datasets consist of either all prokaryotic or all eukaryotic viruses. 'Positive' viruses are those that are reported to be associated with the given host species. A matching number of 'negative' viruses are randomly sampled from all other prokaryotic or eukaryotic viruses. The prokaryote datasets consist of nearly all dsDNA viruses which have on average nearly 100 proteins coded in their genomes, while the eukaryote datasets include many RNA viruses which contain far fewer proteins with on average only roughly 10 protein sequences per virus. The performance of MIL improves with higher numbers of instances in each bag, therefore we need to increase the threshold for the number of viruses in the eukaryotic training datasets to achieve similar performance with MIL. Accordingly, we set a threshold for minimum positive dataset size to 47 and 125 viruses for constructing prokaryotic and eukaryotic binary datasets, respectively. The aim of setting the threshold is to generate a sufficient number of validation samples, with at least 15 and 40 samples for prokaryotic and eukaryotic hosts, respectively. Finally, we generated 15 prokaryotic host datasets and 5 eukaryotic host datasets (see Fig 2 A).

For training the binary classification models, we use two strategies to sample the negative viruses. Strategy 1 is to subsample viruses from those that infect any host from a different genus than those infected by the positive viruses, this means that the negative samples will be more distinct from the positive samples. Strategy 2 is to select from those viruses that are associated with hosts belonging to the same taxonomic groups at the ranks of genus, family, order, class, and phylum as those hosts infected by positive viruses. The aim is to keep the negative samples close to the positive samples and make the binary models more sensitive to clarify the difference between positive and negative samples. The number of viruses related to each host is shown in S1 Table). The largest prokaryote dataset is Mycolicibacterium smegmatis with 838 known viruses, followed by Escherichia coli with approximately half the number of viruses. For the Eukaryotic datasets, Humans have by far the largest number of known virus species (1321) with the next highest being the tomato (Solanum lycopersicum) at 277, see S2 Table.

2.2 EvoMIL achieves high performance for binary virus-host prediction

Embedding vectors for each of the proteins of a virus generated with the protein language model, ESM1b, were used as an instance in a "virus bag" for MIL. These labelled bags were used to train the MIL model using 5-fold cross-validation on 80% of the datasets, then each 5-fold model performance was evaluated on the remaining of 20% test data. Each model is evaluated with a range of metrics: AUC, accuracy, F1 score, sensitivity, specificity, and precision. We evaluated the predictive performance of EvoMIL for binary classification using the datasets generated with both Strategy 1 and Strategy 2 above, training a prediction model for each host.

2.2.1 Prokaryotic and Eukaryotic host performance

The heatmaps of evaluation indices for the prokaryotic and eukaryotic host classifiers are presented in Fig 2 A and B. Here, evaluation indices are calculated based on the

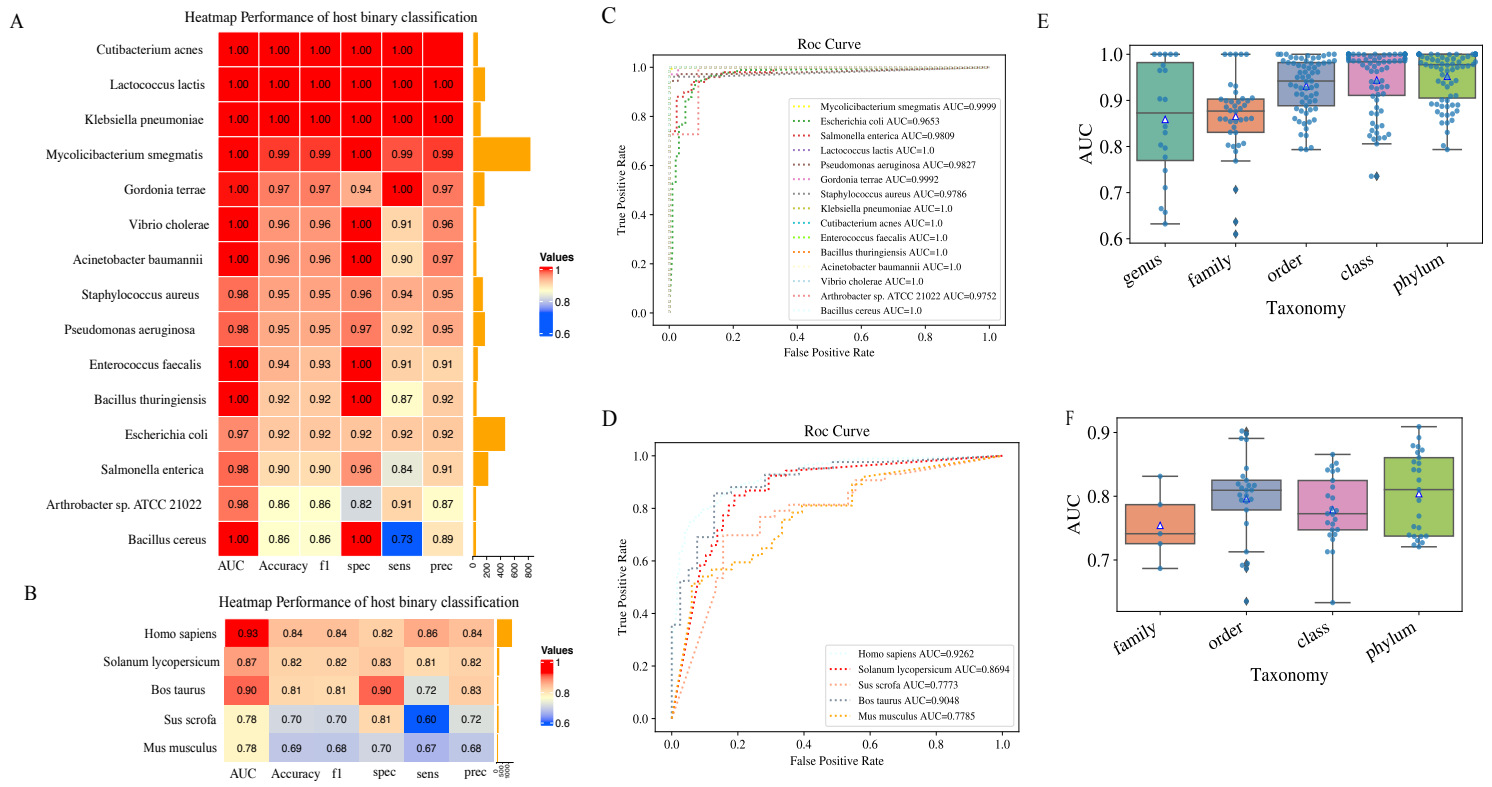

Fig 2. A and B: the heatmap of AUC, accuracy, F1 score, sensitivity, specificity, and precision. There are 15 prokaryotic and 5 eukaryotic host binary classifiers, negative samples are selected by strategy 1; C and D: Roc curves presented 15 prokaryotic hosts and 5 eukaryotic hosts corresponding with heatmap; E and F represent the changes in AUC for prokaryotic and eukaryotic hosts, respectively, when negative samples are selected using strategy 2.

best-performing model with the highest AUC chosen by 5-fold cross-validation. The accuracy is higher than 0.9 except for two hosts, which are 0.86. The ROC curves in Fig 2 C show that all prokaryotic classifiers perform very strongly with each host achieving an AUC value greater than 0.95 and 8 achieving an AUC value of 1.

We also obtained mean and deviation values of evaluation indices for each host, by testing 5 cross-validation models on the host test dataset (see Table S3). EvoMIL shows good performance, with 15/15 hosts achieving a mean AUC value of greater than 0.9 and 8 achieving a mean AUC value of 1. Overall, our results demonstrated that EvoMIL performs well in the binary classification tasks of prokaryotic hosts. More evaluation metrics are included in Table S3.

The ROC curves of the 5 eukaryotic hosts classifiers are presented in Fig 2 D. Here, evaluation indices are calculated by the best-performing model with the highest AUC in 5-fold cross-validation. Although the eukaryote classifiers achieve good performance with AUCs above 0.75, they perform less well than the prokaryote classifiers, with only 3/5 datasets scoring an AUC above 0.85. There may be several explanations for the lower performance. Firstly, the average number of proteins per virus is much lower, resulting in small "bags" for MIL. Secondly, there is a much higher diversity of virus types in the datasets of the eukaryotic hosts, often containing viruses from all Baltimore classes. The virus of these different classes is polyphyletic meaning they will have no common

ancestor and therefore have no shared genes and interact with different host pathways.

The accuracy of each host classifier is shown in Fig 2 B, it is clear that all hosts perform an accuracy higher than 0.8 except two hosts which are roughly 0.7. Homo sapiens obtained the highest accuracy with 0.84; Mus musculus has the lowest accuracy with 0.69.

The mean and deviation values of the evaluation indices for each host are obtained by testing five trained cross-validation models on the test data set (see Table S4). Here, the mean AUC is higher than 0.8 except for the classifiers of two hosts that perform less well, with AUC scores of 0.738±0.14 for Bos taurus and 0.672±0.08 for Canis lupus. Overall, our results demonstrate that EvoMIL is promising in binary classification tasks of eukaryotic hosts.

2.2.2 Sampling negative samples from similar viruses makes binary host classification more challenging

Next, we wanted to test our model with more challenging tasks. Using the second strategy of selecting the negative dataset from viruses related to the positive host at progressively lower taxonomic ranks, from phylum to genus, classification becomes more difficult. Our results show that our PML-MIL models achieve high AUC scores but that distinguishing between viruses of similar hosts is more difficult with a noticeable drop in performance at family and genus levels. In Fig 2 E and F, the box plots show changes in AUC values on prokaryotic and Eukaryotic hosts based on negative selection strategy 2 with five taxonomies genus/family/order/class/phylum. Phylum level (lime colour) presented significantly improves compared with lower taxonomies, especially genus level. Note, at the lower taxonomic ranks there are only sufficient numbers of negative viruses to meet our threshold of 47 for 4 hosts at the genus level, 8 hosts at the family level and 13 hosts at the order level.

2.3 Embedding features outperform protein and DNA k-mer features on Multi-class tasks

To demonstrate that the ESM embeddings are encoding more relevant information than conventional features, we generated feature sets from k-mer composition of the nucleic acid, amino-acid and physio-chemical sequences [10], and implemented ESM-1b and k-mer features on multi-class classification model. Here, we performed multi-class classification extending attention-based MIL by modelling the joint Bernoulli distribution of bag labels. Both prokaryotic and eukaryotic multi-class datasets were constructed using hosts infected by at least 30 viruses. This resulted in 22 classes for the prokaryotes and 36 classes for the eukaryotes. These datasets were used to generate the four different feature sets used in our comparison. Again, we applied 5-fold cross-validation on training datasets, then separately tested trained models on the testing dataset. These models are named ESM-1b, AA_2 and PC_3 and DNA_5, according to different features.

The results for each model are presented in Table 1. Again, we obtain AUC, accuracy and F1 scores by evaluating each of the 5-fold models with the test dataset. Comparing ESM-1b with AA_2, PC_4, and DNA_5, as shown above, for both prokaryotic and eukaryotic hosts, the ESM-1b features have the strongest prediction signal compared with the AA, PC, and DNA k-mers since during testing on prokaryotic hosts, the mean F1 score of EvoMIL is 0.88 which is 8.6%, 12.3%, and 4.1% higher compared with AA_2, PC_4 and DNA_5, respectively; while the counterpart of EvoMIL in eukaryotic hosts is 0.292 which is 0.5%, 1.8% and 3% higher compared with AA_2, PC_4 and DNA_5.(Table 1). Fig 3 A and B respectively show AUC and accuracy of

multi-class MIL by implementing ESM-1b and k-mer features, and AUC and accuracy values are equivalent with Table 1.

Fig 3 A and B respectively show AUC and Accuracy values across ESM-1b and k-mer features on both prokaryotes and eukaryotes. In Fig 3 A, it is clear to see that ESM-1b performs the best AUC and smallest deviation in prokaryotes hosts; on eukaryotes hosts, AUC of ESM-1b shows the smallest deviation and highest mean AUC values compared with k-mer features. Although AA_2 obtained the highest AUC value, it performs the largest deviation. In Fig 3 B, ESM-1b performs the highest accuracy values and the smallest deviation on prokaryotic hosts; on eukaryotes hosts, the highest Accuracy value is obtained by ESM-1b, and ESM-1b performs the smallest deviation except for AA_2. Although AA_2 showed the smallest deviation in accuracy values, ESM-1b obtained the best mean AUC and F1 score (see Table 1).

To evaluate the multi-class MIL models, we collect host datasets associated with a fewer number of viruses, which are between 5 and 30. ESM-1b and k-mer features can be extracted for these datasets, and then test on trained multi-class models. Fig 3 C and D show the accuracy of MIL multi-class models based on ESM-1b and k-mer features (AA_2, PC_4, and DNA_5). Here, if predicted host labels belong to the same genus/family/order/class/Phylum/kingdom with the true host label, we think it is a correct prediction. Then we calculate the percentage of true predictions on all test samples as accuracy values, which are used to evaluate the performance of EMS-1b and k-mere features.

In prokaryotic hosts (Fig 3 C), ESM-1b performs best on the genus level with the highest mean accuracy, 1.8%, while mean accuracy of AA_2, PC_3, DNA_5 are 1.088%, 1.178% and 1.23%, respectively. On top of that, across each taxonomy level, deviation values of ESM-1b are the lowest compared with k-mer features, although ESM-1b does not perform the highest accuracy on family, order, class, and phylum levels. It means that ESM-1b performs best on the genus level, and shows stable accuracy results across all taxonomy levels on prokaryotic hosts. As for eukaryotic hosts (see Fig 3 D), it is clear to see that ESM-1b performs the best accuracy compared with protein and genome k-mer features on each taxonomy level. For example, the mean accuracy percentage of ESM-1b is 36.868% on class level, which are higher by 12.862%, 9.494%, and 11.724% than AA_2, PC_3 and DNA_5. Overall, the multi-class MIL model based on ESM-1b features is able to predict hosts which are associated with less than 30 viruses in eukaryotic datasets. In prokaryotic hosts, we can also find that accuracy values are roughly between 1 and 10 on genus, family and order levels, while accuracy is between 20 and 90 on class, phylum, and kingdom levels. In eukaryotic hosts, accuracy values are roughly between 1 and 6 on genus, family and order levels, while accuracy is between 25 and 50 on class, phylum, and kingdom levels.

Overall, EMS-1b performed best on both prokaryotic and eukaryotic hosts, compared with k-mer features. We also split multi-class datasets into 80% training and 20% test sets, and evaluate the specificity and sensitivity of each host during the multi-class task. In Fig 4, prokaryotic and eukaryotic taxonomy trees combined with specificity and sensitivity heatmap plots are respectively presented in A and B, and the order of the hosts in the heatmap are aligned to the taxonomy tree. Four heatmaps respectively presented the sensitivity and specificity of ESM-1b, AA_2, PC_3 and DNA_5 on each host.

To understand if the host phylogeny explains the prediction performance of multi-class classification, we align the taxonomy tree of hosts with the sensitivity and specificity heatmap Fig 4 and heatmap plots of the number of predicted false hosts on each host Fig 5, aiming to check if false predicted host classes of viruses are more likely to be close hosts in the tree. In prokaryotic hosts, ESM-1b obtained the most number of the highest sensitivity (15/22) and specificity (12/22) compared with protein and DNA

Table 1. The AUC, Accuracy and F1 score of multi-class MIL by using ESM-1b and k-mer features

Host type	Methods	AUC	Accuracy	F1 score
Prokaryotes	ESM-1b	**0.992±0.0**	**0.909±0.0**	**0.88±0.01**
	AA_2	0.979±0.0	0.856±0.01	0.794±0.02
	PC_3	0.969±0.01	0.843±0.01	0.757±0.02
	DNA_5	0.987±0.0	0.882±0.01	0.839±0.01
Eukaryotes	ESM-1b	**0.831±0.01**	**0.494±0.01**	**0.292±0.01**
	AA_2	0.829±0.03	0.494±0.01	0.287±0.01
	PC_3	0.821±0.01	0.479±0.01	0.274±0.02
	DNA_5	0.801±0.02	0.466±0.02	0.262±0.01

Comparison of AUC, accuracy, F1 score between ESM-1b and k-mer feature sets: ESM-1b, AA_2, PC_4 and DNA_5. For each method, training multi-class classification models on 22 prokaryotic hosts and 36 eukaryotic hosts based on 5-fold cross-validation, then mean and deviation values of evaluation metrics are obtained.

k-mer features, while the highest sensitivity and specificity are 21/36 and 14/36 in eukaryotes hosts. In Fig 5, two heatmaps respectively present the number of false predicted hosts which belong to the same taxonomy as the true host in prokaryotes and eukaryotes hosts.

In Fig 5 A, the most number of false predicted hosts are belongs to the family level, here three prokaryotes hosts (Escherichia coli, Klebsiella pneumoniae, Salmonella enterica) account for the most number of false predicted hosts. Besides, Bacillus cereus and Bacillus thuringiensis will predict each other as the host label on the genus level. In Fig 5 B, the most number of false predicted hosts are belongs to the class level, eukaryotes hosts Homo sapiens and Bos taurus account for the most number of false predicted hosts, and they predict each other as the host label on class level, same situation also can be seen between Sus scrofa and Mus musculus.

To understand the relationship between host phylogeny and predicted results, we can take Ralstonia solanacearum as an example in prokaryotes, which performs the lowest sensitivity (see Fig 4 A). Here, 5 false predicted hosts of ESM-1b on Ralstonia solanacearum are 1 Escherichia coli and 4 Pseudomonas aeruginosa, which all belong to the same phylum with Ralstonia solanacearum. In Fig 4 B, it is clear to see that it is more challenging to predict eukaryotic hosts compared with prokaryotes hosts. ESM-1b demonstrates significantly superior performance compared to k-mer features in Nicotiana clevelandii and Cucumis sativus, while its performance on other hosts is comparable. For examples, we can take Homo sapiens as an example, the number of false positive host labels of Homo sapiens based on ESM-1b mainly classified to Pan troglodytes (11), Chlorocebus aethiops (9), Macaca mulatta and Sus scrofa (8). Here, Pan troglodytes belong to the same family; Chlorocebus aethiops and Macaca mulatta belongs to the same order level; Sus scrofa belongs to the same class level.

Overall, viruses are more likely to be misclassified as closely related hosts during multi-class classification, therefore, we assume that close hosts in the taxonomy tree tend to be infected by similar viruses.

2.4 MIL attention weights can be used to interpret which virus proteins are important

Next, we investigate if the proteins important for the prediction can be explained as proteins that are key in virus-host specificity and whether the transformer embeddings are encoding biologically meaningful information about the viral proteins. As we described in the introduction, attention-based MIL learns the weight of each protein in

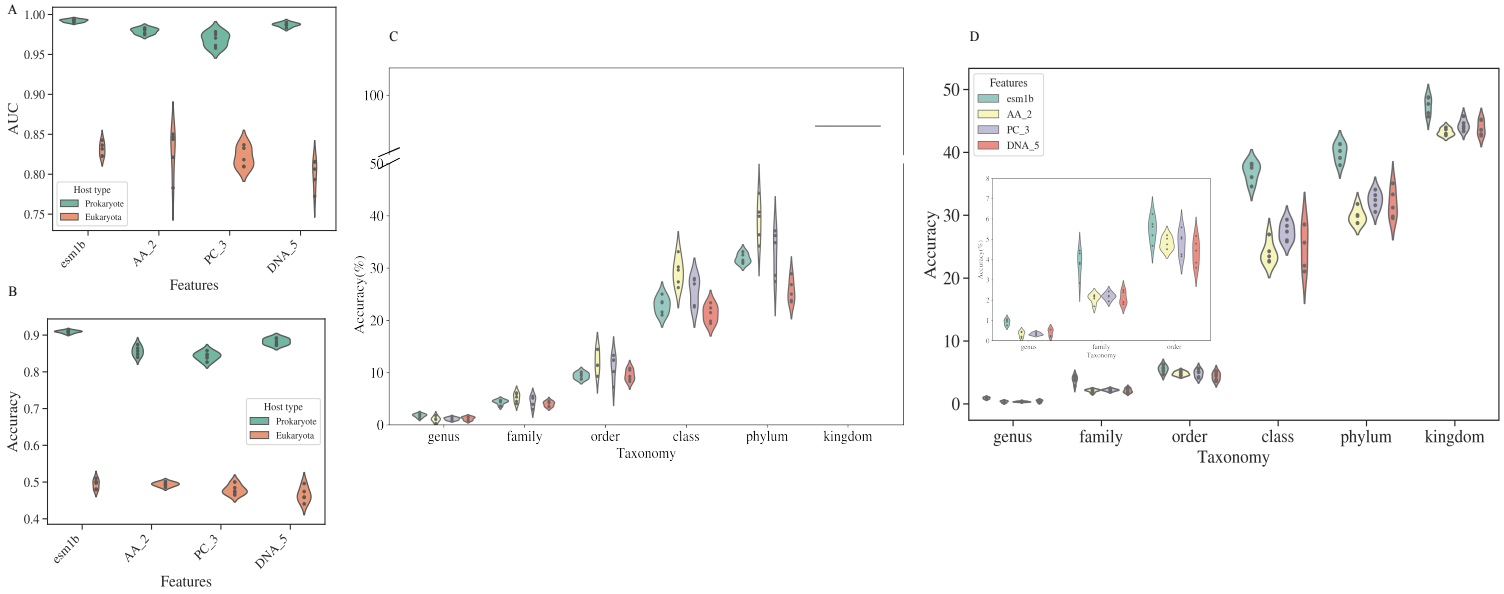

Fig 3. A and B present the AUC and accuracy of prokaryotic and eukaryotic hosts on ESM-1b, AA_2, PC_3 and DNA_5. C and D are the test results of prokaryotic and eukaryotic hosts which are infected by 5 to 30 viruses based on trained 5-fold cross-validation models.

a virus bag, where a high weight indicates that a protein is important for host prediction, and by implication is more likely to be important to virus-host specificity. We used UMAP[11] to investigate whether the embeddings of these important proteins identified by attention-based MIL contain any underlying clustering structures.

For this investigation, we used the viruses of E. coli and H. sapiens which contain the most annotations because they are the most extensively studied hosts from each domain. Using the learned model parameters of the best-performing cross-validation model in the binary classification and multi-class task respectively, we ranked the attention weights of the proteins for the positive viruses and selected the top 5 ranked proteins for each virus. Then we will collect two groups of top proteins embedding based on binary (Fig 6 A and C) and multi-class classification (Fig 6 B and D) models, respectively. These proteins were annotated with functions related to the viral life cycle using GO terms obtained from interproscan [12]. UMAPs were used to visualize any clustering structure that indicates similarities of the protein embeddings (Fig 6 A and B (Escherichia coli) and Fig 6 C and D (Homo sapiens)).

Interestingly both UMAPs show some clustering structure, we assume that these clusters occur because the different functions driving host specificity are captured by the ESM-1b embeddings. In other words, ESM-1b embeddings encode functional and structural signals, and MIL learns some consistent features associated with host specificity.

For the set of top-ranked proteins of the viruses associated with E. coli based on the binary model, only 1.5% (35/2359) were assigned a viral life cycle GO term. Here, 4 kinds of GO terms related to the viral life cycle were used to colour the proteins in UMAP (see Fig. 6 A). This shows that proteins with different GO terms form distinct clusters. More specificity, viral capsid assembly (GO:001906), viral DNA genome packaging (GO:0019073), viral RNA genome replication (GO:0039694), and viral entry into a host cell (GO:0046718) are shown in distinct clusters. In Fig 6 B, there is E. coli

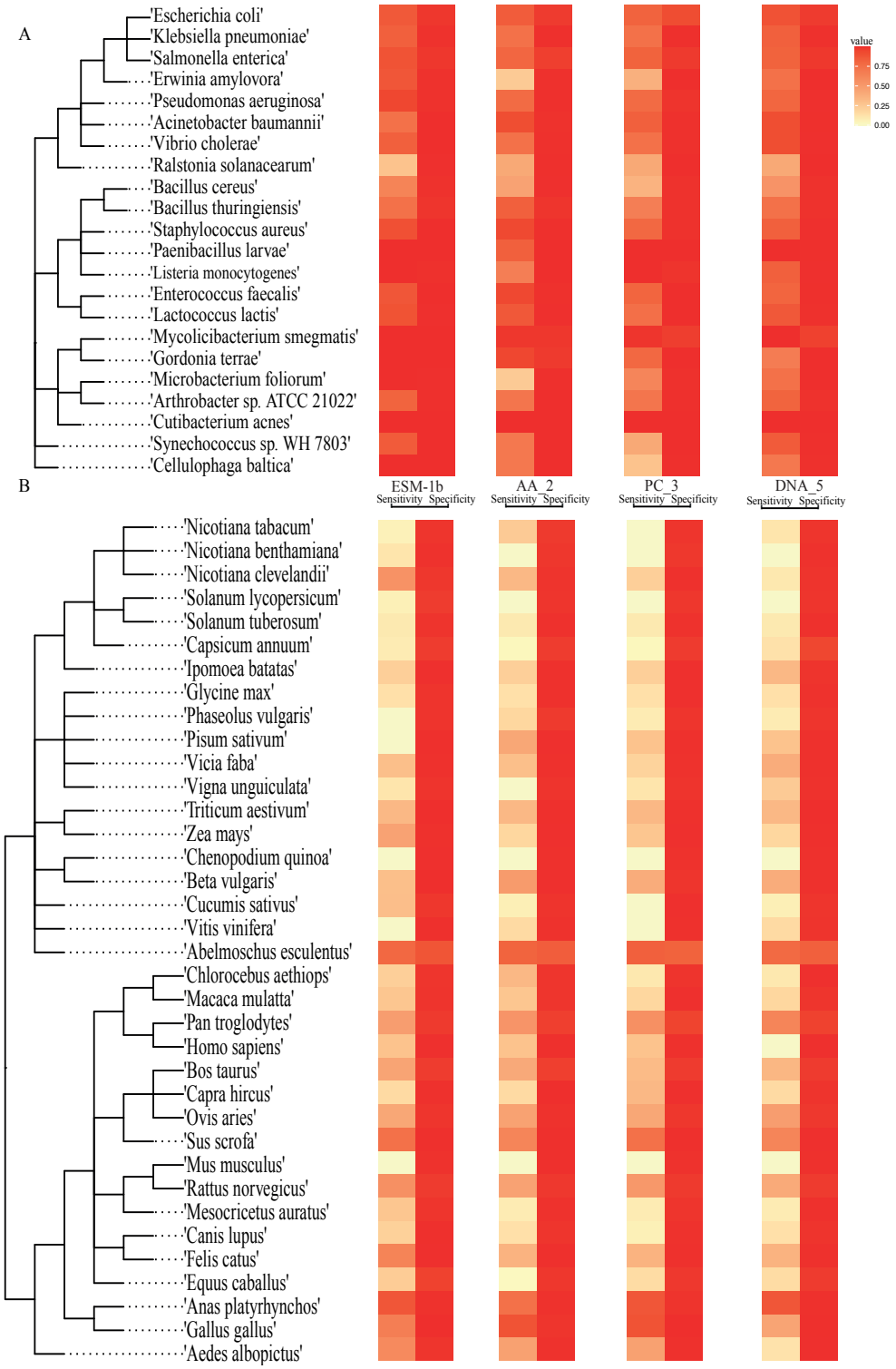

Fig 4. The taxonomy tree and confusion matrix plots of three methods. A and B show the taxonomy tree of 22 prokaryotic and 36 eukaryotic hosts combined with sensitivity and specificity heatmap, respectively, the rest of the subplots aligning with the taxonomy tree are sensitivity and specificity heatmap of ESM-1b, AA_2, PC_3 and DNA_5 separately.

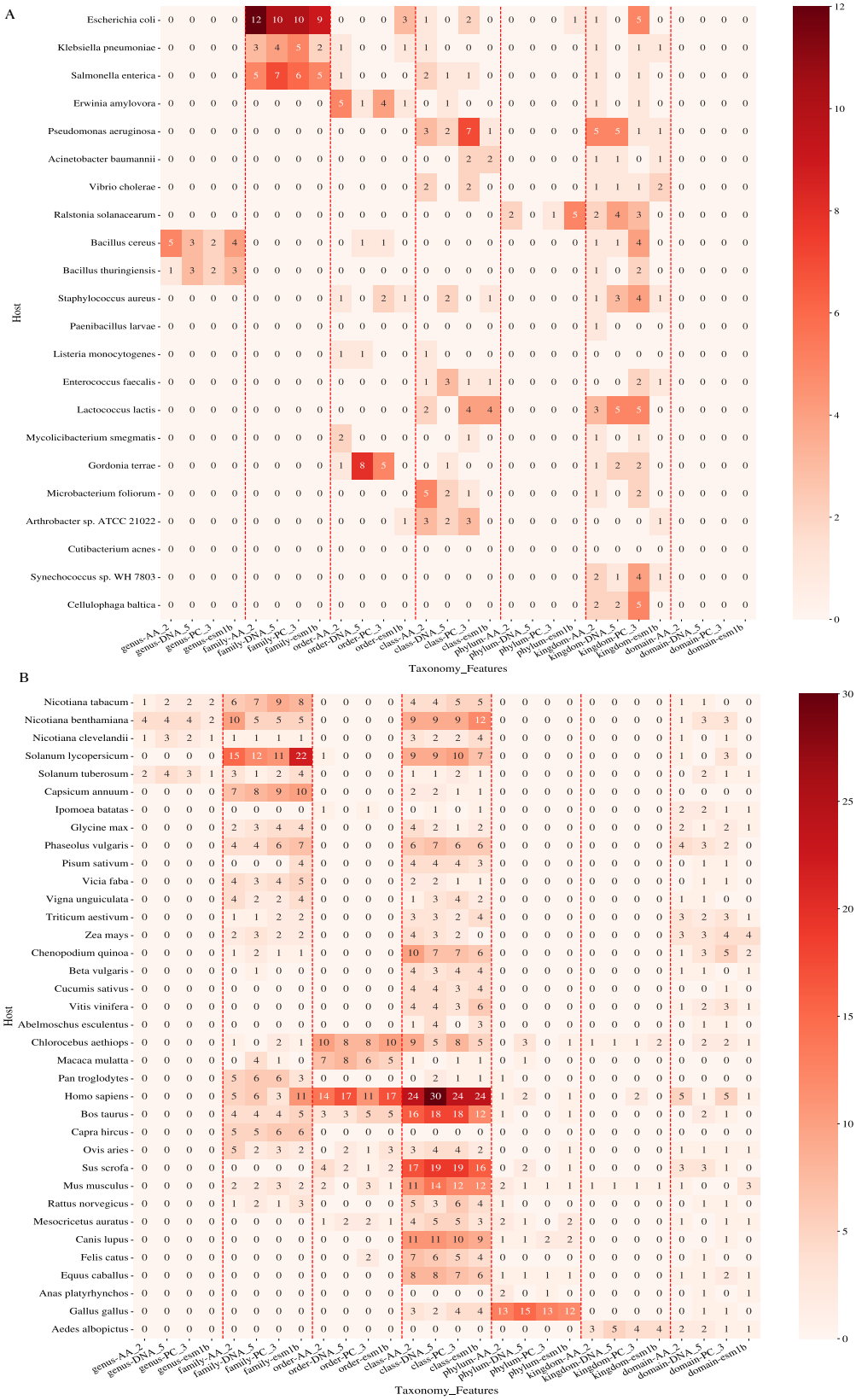

Fig 5. A and B respectively show the heatmap between different features and each host on prokaryotes and eukaryotes, and values are the number of predicted hosts which belong to the same taxonomy with the true host.

UMAP based on the multi-class model, there are more different GO terms, and 3.7% (87/2359) were assigned a viral life cycle GO term. Here, 10 kinds of GO terms related to the viral life cycle were used to colour the proteins in the UMAP. Viral capsid assembly(GO:001906), viral entry into the host cell(GO:0046718), virion assembly(GO:0019068), viral DNA genome packaging(GO:0019073) and viral DNA genome replication(GO:0019073) cluster in the different clusters, in which sub-clusters of viral capsid assembly and virion assembly are two close GO terms. Based on the GO tree, viral capsid assembly is the child node of virion assembly, indicating that the projection of embeddings with high attention weights is explainable for relationships of GO terms.

The UMAP of the embeddings of the top-ranked viral protein associated with Homo sapiens displays a clearer clustering structure (Fig 6 C and D). Overall, the embeddings capture the consistency features of viruses, which are important predictive signals for predicting virus-host associations. In UMAP based on attention weights obtained from binary model (Fig 6 C), 10 different GO terms were retrieved from interproscan, with nearly 23.9% (1280/5357) of the proteins being assigned GO terms. In UMAP based on attention weights obtained from a multi-class model (Fig 6 D), 10 different GO terms were retrieved from interproscan, with roughly 12.6% (676/5357) of the proteins being assigned GO terms. We only choose protein GO terms about the viral life cycle for analysis. Here we marked different colours for each GO term in the UMAP (see Fig 6 C and D). The number of top-5 tanked proteins based on multi-class models is fewer than binary tasks, because

Again, the UMAP shows that different protein functions are clustered into distinct spaces, while GO terms containing similar functions form sub-clusters. For example, in both Fig 6 C and D, viral life cycle (GO:0019058), viral DNA genome packaging (GO:0019073), viral RNA genome replication(GO:0039694), and viral genome replication(GO:0019079) are taking the largest part of percentages and all presented in different clusters, where viral life cycle (GO:0019058) shows in diverse clusters as it is the parent node including a wide variety of GO terms. Several embeddings annotated by viral genome replication (GO:0019079) and viral RNA genome replication (GO:0039694) are tightly clustered in the close space, and viral RNA genome replication is part of viral genome replication based on the GO tree, which indicates that ESM-1b can capture protein functions during host prediction. Binary and multi-class shared the most of GO terms, but virion assembly (GO:0019068) is only shown in Fig 6 C which is based on binary models.

To visualize the projection of UMAP in proteins, we choose the top-5 ranked proteins to present protein GO annotation clusters. In Fig 7, we can see that top-5 ranked proteins annotated by GO terms present important proteins with high weights, and all proteins contain many zero weights, so selecting the top-5 ranked proteins allows us to collect proteins with Go annotation and high weights, as the aim is to present key proteins which are assigned high weights binary and multi-class models. We mark the GO annotation for proteins and select two GO annotations as examples. For example, in Fig 7 A, it is clear to see that the position of GO annotation of top-5 ranked proteins is the same, meaning that we can keep GO annotations although selecting top-5 ranked proteins.

Finally, we looked at the prediction and attention weights of SARS-CoV-2 [13] (Severe acute respiratory syndrome coronavirus 2) using the H.sapiens binary classifier from section 2.2.1. The original Wuhan strain with NCBI Reference Sequence NC_045512.2 was predicted to be a human virus with a probability of 0.432 compared to the SVM-kmer classifiers [10] of less than 0.01. The top three ranked proteins are the Non-structural proteins sub-sequences, Nucleocapsid protein and spike sub-sequences (1022 and 251 amino acids long), with weights of 0.622, 0.093 and 0.066 respectively.

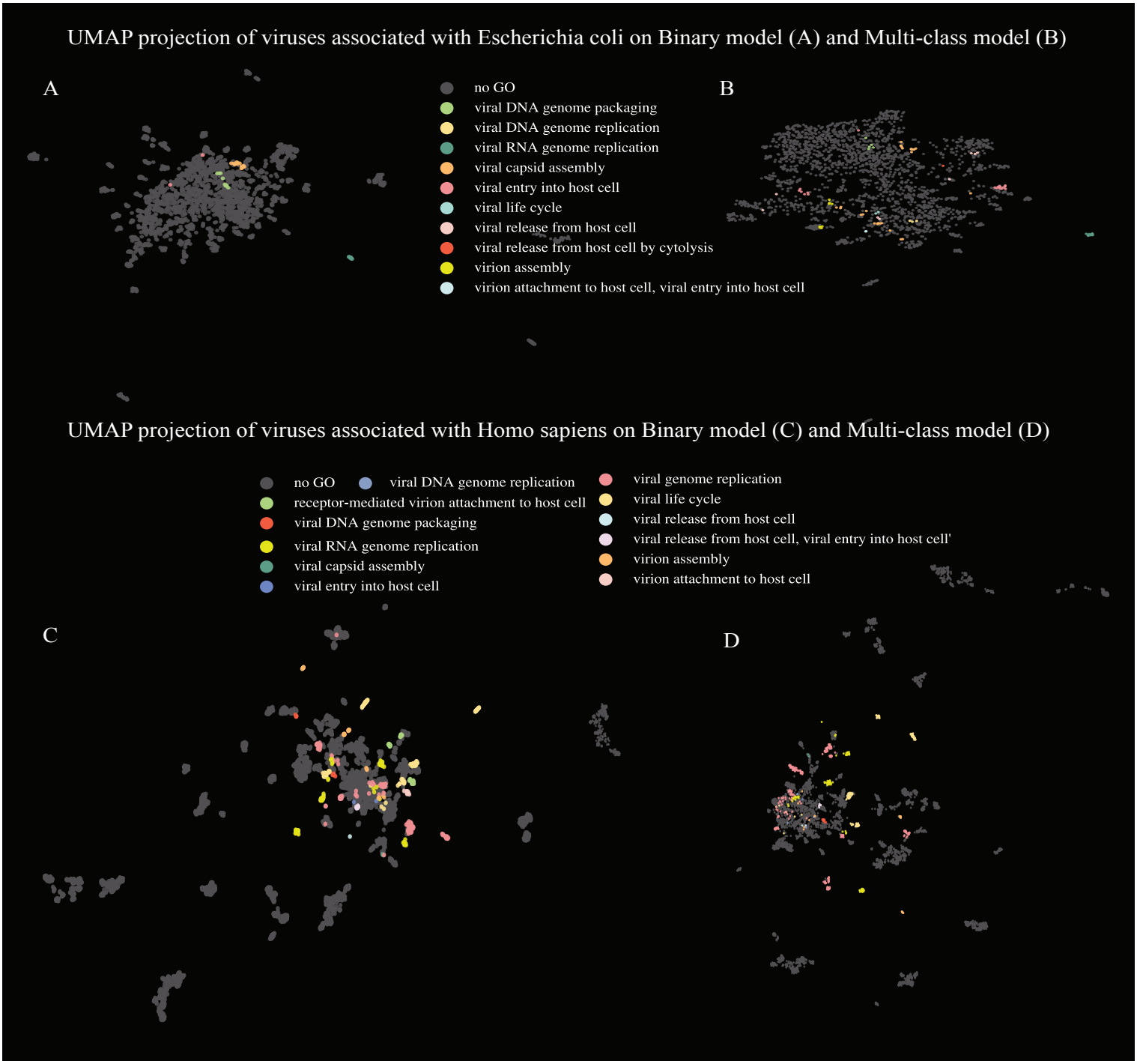

Fig 6. UMAPs projection of protein embeddings of viruses associated with Escherichia coli and Homo sapiens. Figure A and B show the two-dimensional distribution of top-5 ranked protein embeddings of each virus which is associated with Escherichia coli; figure C and B show the distribution of embedding vectors of top-5 ranked proteins of each virus which is associated with Homo sapiens. Here, the protein weights in Figures A and C are obtained by binary models, while Figures B and D are obtained by multi-class classification models. Protein GO terms regarding the viral life cycle of viruses are marked by different colours for understanding predictive signals of proteins captured by the pre-trained transformer model.

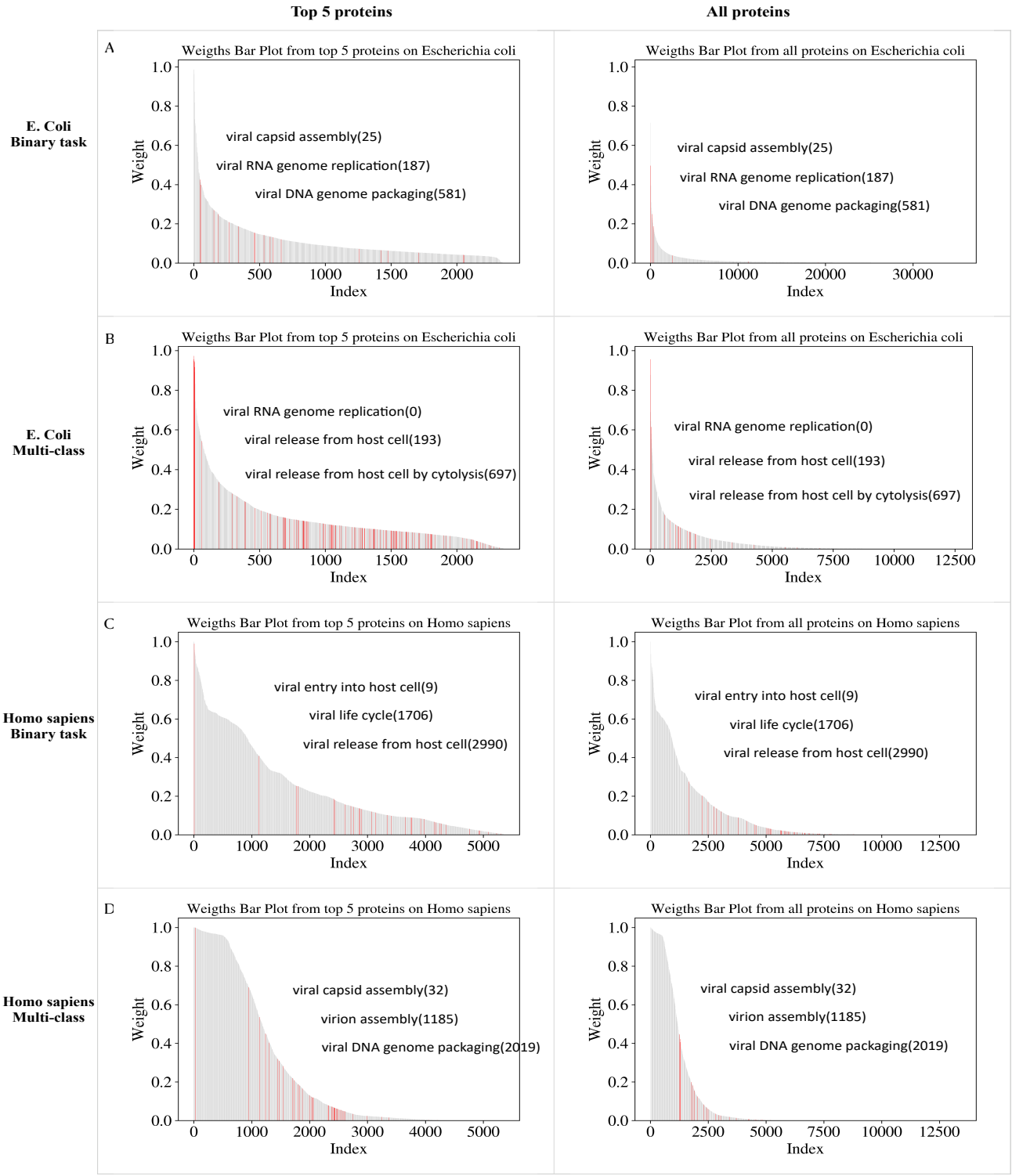

Fig 7. The bar plots of top 5 protein weights and all protein weights Figure A and B show the bar plot of protein weights obtained by binary classification models for Escherichia coli and Homo sapiens separately, while Figure C and D show the bar plot of protein weights obtained by multi-class classification models from Escherichia coli and Homo sapiens separately. Here, the left subplot present top-5 ranked protein weights and the right subplot presents all protein weights.

The spike protein contains the receptor binding site which is responsible for cell entry. Nucleocapsid protein is an RNA binding protein that plays a critical role at many stages of the viral life cycle making direct interactions with many host proteins [14]. More N protein and host proteins interactions can be retrieved from protein-protein interaction databases in UniPort (`https://www.uniprot.org/uniprotkb/P0DTC9/entry#interaction`). The viral process GO terms of these top-ranked proteins are "fusion of virus membrane with the host plasma membrane" (GO:0019064), "fusion of virus membrane with host endosome membrane" (GO:0039654), "receptor-mediated virion attachment to host cell" (GO:0046813) and "endocytosis involved in viral entry into host cell" (GO:0075509). It is known that these GO terms are associated with viral infections, indicating that protein attention weights obtained by EvoMIL perform potentials in explaining important proteins contributing to virus-host associations.

3 Discussion

In this paper, we introduce EvoMIL, a novel method for virus-host prediction. Inspired by the success of NLP approaches in biology, we demonstrate the power of using the protein transformer language model ESM-1b to generate embeddings of viral proteins that are highly effective features for virus-host classification tasks. ESM-1b is capturing meaningful biological/host information from viral proteins despite the training dataset being mainly proteins from cellular life with only 1% being viral proteins. We demonstrate that attention-based MIL can identify which of a virus's proteins are most important for prediction, and by implication which proteins may be key to virus-host specificity.

Our classification results show that EvoMIL is able to predict the host at the species level with high AUC, accuracy and F1 scores. The prokaryote classifiers achieved a mean AUC of 0.943 whereas the eukaryotic classifiers all achieved a mean AUC of 0.857 (see Table S3 and Table S4). Eukaryote host prediction is a more difficult task for two reasons: viruses associated with eukaryotic hosts are much more diverse S1 Table with examples from across all seven Baltimore classes compared to the prokaryote viruses which are mainly double-stranded DNA viruses; secondly, the eukaryote datasets contain many RNA viruses which have far fewer proteins, this makes it more challenging for MIL which needs many instances in each bag to perform well. We showed that ESM-1b protein features outperform more conventional amino acid and physio-chemical property k-mer features on binary classification tasks, achieving a mean accuracy of 0.95 for all 21 prokaryote hosts, an improvement of 3.66% and 1.47% over AA_2 and PC_4 features. Furthermore, the clustering structure seen in the UMAP projections (Fig 6) indicates that embeddings are able to capture functionally related information of the high-ranked proteins associated with host specificity.

Transformers enable a one-step pipeline to generate dense feature vectors from viral genomes and learn multiple layers of information about the structural and functional properties of proteins in an unbiased way. Multiple-instance learning is able to capitalize on any host signal contained within individual protein sequences, circumnavigating the need to represent each virus with a single feature vector. This enables it to find common patterns/signals across bags of proteins that differ greatly in both size and content. An additional advantage of attention-based MIL is the ability to identify which proteins are important in prediction, further analysis is needed to test whether this can be exploited to uncover the mechanisms behind virus-host specificity.

In this paper we took a virus-based approach in which only information from the viral genomes is used to generate features, this limits us to classifying viruses for those hosts that have a minimum number of known viruses. Also, by limiting the viral

sequences to only Refseq sequences to reduce redundancy, we have further narrowed the range of hosts we can predict. The ultimate host predictor would be able to make predictions for hosts which have no or few known viruses, to do this we will include host information. Combining virus and host-based approaches has been shown to greatly increase the host range of a prediction tool whilst maintaining a low false discovery rate, Roux et al.[15] and Wang et al. [16]. In the future, we will make use of the much larger numbers of host-annotated virus genomes available in public databases and include host information to construct a model that can make predictions for associations for any pair of viruses and host.

Materials and Methods

Data

The current study was done using datasets that were collected from the Virus-Host database VHDB, (`https://www.genome.jp/virushostdb/`) [9]. The VHDB contains a manually curated set of known species-level virus-host associations collated from a variety of sources, including public databases such as RefSeq, GenBank, UniProt, and ViralZone and evidence from the literature surveys. For each known interaction, this database provides NCBI taxonomic ID for the virus and host and the Refseq IDs for the virus genomes. We download datasets on 20-10-2021, along with the associated Fasta files containing the raw viral genomes and the FAA file with translated CDS for each viral protein. At this point, the VHDB contained 17733 associations between 12650 viruses and 3740 hosts that were used to construct binary datasets for both prokaryotic and eukaryotic hosts.

Constructing balanced binary dataset

In this study, we only collect the Refseq genome [17] for each virus. The aim is to reduce the amount of redundant or similar sequences in the datasets. We obtain known virus-host associations for prokaryotic and eukaryotic hosts from the Virus-Host database VHDB [9], and remove viruses whose protein sequences do not exist in the Refseq database.

Balanced binary datasets were constructed for each host with an equal number of positive and negative virus sets to obtain a balanced data set for binary classification tasks. For both the prokaryotic and eukaryotic datasets, we collected 4696 associations between 4696 viruses and 498 prokaryotic hosts at the species level; 9595 positive associations from 9595 viruses and 1665 eukaryotic hosts at the species level.

For each binary virus-host association data set, viruses can be represented by $V = \{V_1, V_2, ..V_V\}$, and the host is represented by H. The positive labels are consisted of known associations from VHDB [9], and the set of positive viruses can be represented by $V_{pos} = \{V_1, V_2, ..V_P\}$, where $P = \frac{V}{2}$. As we know, there is no information on negative associations in the public datasets. Most phages are likely to infect a range of hosts belonging to the same species[18, 19], therefore, in order to mitigate possible errors, some researchers [20, 21] construct negative virus-host associations by selecting those viruses from the remaining viruses that do not infect the given host, instead of relating to hosts with different taxa from the given host. Inspired by this method, we construct putative negative virus-host associations by collecting viruses from a specific range. As for a given host H in the prokaryotic datasets, all viruses are represented by $V_{pro} = \{V_1, V_2, ..V_T\}$, we collect negative viruses $V_{neg} = V_{pro} - V_{pos}$, which are not related with the given host; and host species taxonomy of V_{neg} are different from host H, on top of that, in order to get the balance datasets for model training and test, the

size of datasets V_{neg} is equal to V_{pos}, we randomly choose viruses from putative negative virus set V_{neg}, finally we get the same number of positive and negative virus datasets, it can be represented by $V_{neg} = \{V_{P+1}, V_{P+2}, ..V_V\}$. The number of viruses related to the given host is at least equal to 30 and 45 for the prokaryotic and eukaryotic datasets, so the binary host data set contains at least 60 and 90viruses. In addition, there exist some segment and non-segment viruses; segment viruses mean that these viruses have multiple RefSeq sequences and need to be combined to represent complete viral sequences. Non-segment viruses mean that viruses are not split into multiple sequences, we can only select one RefSeq sequence, although those have multiple virus reference accession numbers. After setting the threshold for the number of viruses at the species level, combining segment viruses and removing redundancy non-segment viruses, we get 21 prokaryotic datasets and 12 eukaryotic datasets.

Feature extraction

The Pre-trained ESM-1b is used to transform protein sequences into fixed-length embedding vectors that are then as features for downstream tasks such as binary and multi-class classification. For n protein sequences $(h_1, ..., h_n)$ input into ESM-1b [8], we obtain n embedding vectors $\mathbf{X} = (\mathbf{x_1}, ...\mathbf{x_n})$ each of dimensions 1024. ESM-1b was run with the following parameters: learning rate is 0.0005, and weight decay is 10e-5. In the pre-training process, the representation is projected to log probabilities $(y_1, ..., y_n)$, the model posterior of amino acid at position i is represented by a softmax over y_i, the output embeddings $(h_1, ..., h_n)$ can be applied in downstream tasks, (the parameters are "–repr_layers 33 –include mean per_tok"). Here, we adapt virus embedding vectors to the supervised Attention-based MIL learning tasks.

There are two special cases of sequences that need to be pre-processed to be suitable as input for ESM-1b:

1. In NCBI, some protein sequences include the amino acid J, which is used to refer to unresolved leucine (L) or isoleucine (I) residues. However, the ESM-1b model does not include the token 'J' and they must be removed before processing. In order to remove J from NCBI sequences, we randomly replace J with either Leucine(L) or Isoleucine(I).

2. The ESM-1b model can only process sequences with a maximum of 1024 tokens including the beginning-of-sequence token, BOS, and the ending-of-sequence token, EOS, meaning that the maximum length of the protein sequence is 1022 amino acids. The parameter '–truncation' can be used to crop a longer sequence, but this will result in a loss of some part of the sequence information. In order to include information from the entire protein sequence, we split longer sequences and process the sections simultaneously. For protein sequences longer than 1022, we split the sequence into sections of 1022 amino acids, if the final section is shorter than 25 amino acids it is discarded as it is considered too short to include meaningful information. So that for a given protein sequence H, of length $len(H)$:

   if $len(H) < 1022 + 25$ then truncate the sequence

   else if $len(H) \geq 1022 + 25$ split the sequence.

   For example, if we have a sequence of length 2049 $(2 \times 1022 + 25)$, it will be cut into two sub-sequences with the length of 1022, and one sub-sequence with the length of 25. Resulting in three embedding vectors for the single protein sequence that are all assigned to the same label as the parent protein and will be considered instances in the attention-based MIL.

Attention-based Multiple instance learning

Multiple instance learning

We represent a virus as a set of protein embeddings $X = \{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, ..., \mathbf{x}_M\}$. The label of the virus is $Y \in \{0, 1\}$. If $Y = 1$, the virus is known to be associated with the host, otherwise, the virus is not associated with the host. But individual label $\{y_1, y_2, y_3, ..., y_M\}$ is unknown. Multiple instance learning (MIL) is used to predict a label for a bag with a set of instances, so it was used to predict a host label for a set of proteins from a virus. The assumption about the label Y can be represented as follows:

$$Y = \begin{cases} 0, & \text{iff} \sum_m y_m = 0, \\ 1, & \text{otherwise.} \end{cases} \tag{1}$$

The MIL model can be interpreted as a probabilistic model in which the label of the bag is distributed according to the Bernoulli distribution with the parameter $\theta(X) \in \{0, 1\}$, which is invariant to the permutation of instances.

$$p(Y|X) = \theta(X)^Y (1 - \theta(X))^{1-Y}, \tag{2}$$

In the multi-class problem, the model can be trained with the softmax loss instead.

Given an input instance x_m, the whole scoring function of the MIL can be represented as Eq (3) [22]:

$$f(X) = g(\sum_{x_m \in X} [\varphi(x_m)]), \tag{3}$$

Where, $f(X)$ is a symmetric function, the orders of the instances do not affect the results of the scoring function $f(X)$. Here g and φ are suitable transformations, and the two transformations are modelled using neural networks. Generally, the algorithm can be implemented by modelling a permutation-invariant function $\theta(X)$. (i) φ: generate transformed instances; (ii) $f(X)$: combine transformed instances; (iii) g: obtain the final label based on the combined instances.

MIL is the embedding level, instances are mapped to a low-dimensional embedding based on the function φ, and then the bag representation can be obtained by MIL pooling. Here the bag representation is not related to the number of instances in the bag, finally, a bag-level classifier was applied to process the bag representation and generate $\theta(X)$.

Attention-based pooling

To explain protein embeddings of viruses and learn each feature of protein sequences of viruses, an attention mechanism is considered as MIL pooling [7]. This method applied a weighted average of instances, each instance's weight is calculated by the neural network, and the sum of weights is 1. Here $U = \{u_1, u_2, u_3, ..., u_M\}$ is a bag of M embedding vectors, the MIL pooling is represented by Eq (4):

$$\boldsymbol{z} = \sum_{m=1}^{M} a_m \boldsymbol{u}_m, \tag{4}$$

Here, the attention mechanisms can generate a weight for each protein of viruses, and it can be represented as:

$$a_m = \frac{\exp\{\boldsymbol{w}^\top \tanh(\boldsymbol{V}\boldsymbol{u}_{\mathrm{m}}^\top)\}}{\sum_{j=1}^{M} \exp\{\boldsymbol{w}^\top \tanh(\boldsymbol{V}\boldsymbol{u}_j^\top)\}}, \tag{5}$$

Here, a_m is the weight for the corresponding protein sequence, $w \in R^{L\times 1}$ and $V \in R^{L\times M}$ are parameters, $tanh(.)$ is the hyperbolic tangent function [7] (This code is available from `https://github.com/AMLab-Amsterdam/AttentionDeepMIL?utm_source=catalyzex.com`). We use these weights to quantify the importance of proteins to host prediction.

Evaluation

To evaluate our classifiers we use six evaluation metrics, including AUC, accuracy, F1 score, sensitivity, specificity, and precision as defined below. The two main evaluation metrics used in our analysis and explanation are AUC and accuracy. The area under the receiver operating characteristic curve (AUC) is often used to evaluate machine learning performance; accuracy is the ratio of the number of correctly predicted samples to the total number of samples.

True positive (TP), true negative (TN), false positive (FP), and false negative (FN) are the parameters used to calculate specificity, sensitivity, and accuracy. True positive(TP): predicted label and known label are positive. True Negative(TN): predicted label and known label are negative. False Positive (FP): The predicted label is positive, which is negative. False Negative(FN): The predicted label is negative, which is positive. Sensitivity is the percentage of positive samples; specificity is the percentage of negative samples. The F1 score is the harmonic mean of precision and recall. Here, we set average='macro' to calculate the F1 score and precision for each label, and get their unweighted mean. The formula of the evaluation indices is as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}, \tag{6}$$

$$Sensitivity = \frac{TP}{TP + FN}, \tag{7}$$

$$Specificity = \frac{TN}{TN + FP}, \tag{8}$$

$$Precision = \frac{TP}{TP + FP}, \tag{9}$$

$$Recall = \frac{TP}{TP + FN}, \tag{10}$$

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall}, \tag{11}$$

Acknowledgments

This research was funded by European Union's Horizon 2020 research and innovation program, under the Marie Skłodowska-Curie Actions Innovative Training Networks grant agreement no. 955974 (VIROINF). The authors also acknowledge funding from the Medical Research Council (MRC, $\mathrm{MC}_U U_1 2014/12), the Biotechnology and Biological Sciences Research Council (BBSRC, BB/V01$

S1 Appendix

Code and trained models are available on `https://github.com/liudan111/EvoMIL`.

S1 Table. Results of 21 prokaryotic hosts. The 21 Prokaryotic datasets, show the species name, the number of positive viruses associated with the host for each dataset, and the ANI values of viruses in each host.

S2 Table. Results of 12 eukaryotic hosts. The 12 Eukaryote hosts datasets, show the species name, the number of positive viruses associated with the host for each dataset, and ANI values of viruses in each host.

S3 Table. Table of GO ID and GO terms. For understanding the protein functions of viruses, retrieving the GO term for each GO ID is presented in the UMAP plots.

References

1. Roux S, Adriaenssens EM, Dutilh BE, Koonin EV, Kropinski AM, Krupovic M, et al. Minimum information about an uncultivated virus genome (MIUViG). Nature biotechnology. 2019;37(1):29–37.
2. Roux S, Enault F, Hurwitz BL, Sullivan MB. VirSorter: mining viral signal from microbial genomic data. PeerJ. 2015;3:e985.
3. Staals RH, Brouns SJ. Distribution and mechanism of the type I CRISPR-Cas systems. In: CRISPR-Cas systems. Springer; 2013. p. 145–169.
4. Horvath P, Barrangou R. CRISPR/Cas, the immune system of bacteria and archaea. Science. 2010;327(5962):167–170.
5. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. Advances in neural information processing systems. 2017;30.
6. Maron O, Lozano-Pérez T. A framework for multiple-instance learning. Advances in neural information processing systems. 1997;10.
7. Ilse M, Tomczak J, Welling M. Attention-based deep multiple instance learning. In: International conference on machine learning. PMLR; 2018. p. 2127–2136.
8. 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. Proc Natl Acad Sci U S A. 2021;118(15).
9. Mihara T, Nishimura Y, Shimizu Y, Nishiyama H, Yoshikawa G, Uehara H, et al.. Linking Virus Genomes with Host Taxonomy; 2016.
10. Young F, Rogers S, Robertson DL. Predicting host taxonomic information from viral genomes: A comparison of feature representations. PLoS Comput Biol. 2020;16(5):e1007894.
11. McInnes L, Healy J, Melville J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. ArXiv e-prints. 2018;.

12. Jones P, Binns D, Chang HY, Fraser M, Li W, McAnulla C, et al. InterProScan 5: genome-scale protein function classification. Bioinformatics. 2014;30(9):1236–1240.

13. Wu F, Zhao S, Yu B, Chen YM, Wang W, Song ZG, et al. A new coronavirus associated with human respiratory disease in China. Nature. 2020;579(7798):265–269.

14. Caruso IP, dos Santos Almeida V, do Amaral MJ, de Andrade GC, de Araújo GR, de Araújo TS, et al. Insights into the specificity for the interaction of the promiscuous SARS-CoV-2 nucleocapsid protein N-terminal domain with deoxyribonucleic acids. International journal of biological macromolecules. 2022;203:466–480.

15. Roux S, Camargo AP, Coutinho FH, Dabdoub SM, Dutilh BE, Nayfach S, et al. iPHoP: an integrated machine-learning framework to maximize host prediction for metagenome-assembled virus genomes. bioRxiv. 2022;.

16. Wang W, Ren J, Tang K, Dart E, Ignacio-Espinoza JC, Fuhrman JA, et al. A network-based integrated framework for predicting virus–prokaryote interactions. NAR genomics and bioinformatics. 2020;2(2):lqaa044.

17. Tatusova T, Ciufo S, Fedorov B, O'Neill K, Tolstoy I. RefSeq microbial genomes database: new representation and annotation strategy. Nucleic acids research. 2014;42(D1):D553–D559.

18. Flores CO, Meyer JR, Valverde S, Farr L, Weitz JS. Statistical structure of host–phage interactions. Proceedings of the National Academy of Sciences. 2011;108(28):E288–E297.

19. Ben-Hur A, Noble WS. Choosing negative examples for the prediction of protein-protein interactions. In: BMC bioinformatics. vol. 7. Springer; 2006. p. 1–6.

20. Leite DMC, Brochet X, Resch G, Que YA, Neves A, Peña-Reyes C. Computational prediction of inter-species relationships through omics data analysis and machine learning. BMC bioinformatics. 2018;19(14):151–159.

21. López JF, Sotelo JAL, Leite D, Peña-Reyes C. Applying one-class learning algorithms to predict phage-bacteria interactions. In: 2019 IEEE Latin American Conference on Computational Intelligence (LA-CCI); 2019. p. 1–6.

22. Salakhutdinov SRBPR, Zaheer AJSM, Kottur S. Deep Sets. Advances in Neural Information Processing (NIPS). 2017;.

Table S1. Prokaryotic hosts: hostname, the number of viruses associated with the host

Index	Host name	Number of viruses	Mean ANI	Mean_count_protein
1	Mycolicibacterium smegmatis	838	29.1	105
2	Escherichia coli	474	10.8	130
3	Salmonella enterica	231	10.8	115
4	Lactococcus lactis	185	43	56
5	Pseudomonas aeruginosa	182	10.5	91
6	Gordonia terrae	176	31.4	96
7	Staphylococcus aureus	152	59.2	86
8	Klebsiella pneumoniae	121	13.4	94
9	Cutibacterium acnes	81	92.2	45
10	Enterococcus faecalis	80	29.4	85
11	Bacillus thuringiensis	62	28.2	171
12	Acinetobacter baumannii	60	24.4	111
13	Vibrio cholerae	59	13.7	80
14	Arthrobacter sp. ATCC 21022	55	23.2	76
15	Bacillus cereus	54	23.8	133

The 21 Prokaryotic datasets, showing the species name, the number of viruses associated with the host for each dataset, ANI values of viruses in each prokaryotic dataset, and the average number of protein sequences of viruses associated with the host.

Table S2. Eukaryotic hosts: hostname and the number of viruses associated with the host

Index	Host name	Number of viruses	Mean ANI	Mean_count_protein
1	Homo sapiens	1321	6.7	7
2	Solanum lycopersicum	277	27.9	4
3	Sus scrofa	219	3.7	10
4	Bos taurus	202	2.9	9
5	Mus musculus	174	12.6	16

The 12 Eukaryote hosts datasets, showing the species name, the number of viruses associated with the host for each dataset, ANI values of viruses in each Eukaryote dataset, and the average number of protein sequences of viruses associated with the host.

Table S3. Results for the prokaryotic hosts

Host name	AUC	Accuracy	f1	specificity	sensitivity	precision
Mycolicibacterium smegmatis	0.999±0.0	0.993±0.0	0.993±0.0	0.998±0.0	0.99±0.0	0.993±0.0
Escherichia coli	0.947±0.01	0.893±0.02	0.892±0.02	0.868±0.07	0.915±0.03	0.896±0.02
Salmonella enterica	0.965±0.03	0.899±0.02	0.897±0.02	0.94±0.02	0.851±0.06	0.904±0.01
Lactococcus lactis	0.999±0.0	0.995±0.01	0.995±0.01	1.0±0.0	0.988±0.02	0.995±0.01
Pseudomonas aeruginosa	0.978±0.01	0.953±0.01	0.953±0.01	0.956±0.02	0.951±0.02	0.954±0.01
Gordonia terrae	0.998±0.0	0.975±0.01	0.975±0.01	0.966±0.01	0.983±0.02	0.975±0.01
Staphylococcus aureus	0.974±0.01	0.964±0.01	0.963±0.01	0.962±0.0	0.966±0.02	0.963±0.02
Klebsiella pneumoniae	0.986±0.02	0.963±0.03	0.963±0.03	0.978±0.05	0.945±0.04	0.966±0.03
Cutibacterium acnes	0.68±0.44	0.842±0.22	0.751±0.34	0.6±0.55	1.0±0.0	0.721±0.38
Enterococcus faecalis	0.995±0.0	0.944±0.01	0.934±0.02	1.0±0.0	0.922±0.02	0.917±0.02
Bacillus thuringiensis	0.992±0.01	0.928±0.02	0.927±0.02	1.0±0.0	0.88±0.03	0.924±0.02
Acinetobacter baumannii	0.984±0.02	0.933±0.02	0.929±0.02	1.0±0.0	0.84±0.05	0.949±0.02
Vibrio cholerae	0.957±0.03	0.925±0.03	0.924±0.04	0.969±0.04	0.873±0.05	0.931±0.04
Arthrobacter sp. ATCC 21022	0.953±0.02	0.873±0.05	0.873±0.05	0.873±0.08	0.873±0.05	0.875±0.05
Bacillus cereus	0.977±0.03	0.864±0.03	0.862±0.03	0.982±0.04	0.745±0.04	0.886±0.03

The performance of the binary classifiers for prokaryotic hosts. Here we show evaluation indices by testing 5-fold cross-validation models on each host dataset, then calculate mean and deviation values. The evaluation metrics include AUC, accuracy, f1, specificity, sensitivity, and precision.

Table S4. Results for the eukaryotic hosts

Host name	AUC	Accuracy	f1	specificity	sensitivity	precision
Homo sapiens	0.918±0.01	0.84±0.01	0.84±0.01	0.856±0.03	0.823±0.04	0.842±0.01
Solanum lycopersicum	0.851±0.01	0.8±0.03	0.8±0.03	0.808±0.06	0.793±0.03	0.801±0.03
Sus scrofa	0.761±0.01	0.727±0.02	0.726±0.02	0.791±0.04	0.667±0.04	0.733±0.02
Bos taurus	0.898±0.0	0.837±0.02	0.835±0.02	0.895±0.03	0.774±0.07	0.844±0.01
Mus musculus	0.762±0.02	0.697±0.01	0.696±0.01	0.703±0.02	0.691±0.03	0.697±0.01

The performance of 5 binary classifiers of 5 eukaryotic hosts. Here we show evaluation indices by testing 5-fold cross-validation models on each host dataset, then calculate mean and deviation values. The evaluation metrics include AUC, accuracy, f1, specificity, sensitivity, and precision.

Table S5. Table of GO ID and GO terms

GO ID	GO terms
GO:0019073	viral DNA genome packaging
GO:0039666	virion attachment to host cell pilus
GO:0039694	viral RNA genome replication
GO:0019069	viral capsid assembly
GO:0046797	viral procapsid maturation
GO:0019068	virion assembly
GO:0019079	viral genome replication
GO:0019062	virion attachment to host cell
GO:0039664	lysis of host organelle involved in viral entry into host cell
GO:0019068, GO:0046760	virion assembly, viral budding from Golgi membrane
GO:0019058	viral life cycle
GO:0046813, GO:0075509	receptor-mediated virion attachment to host cell, endocytosis involved in viral entry into host cell
GO:0046718	viral entry into host cell
GO:0075509	endocytosis involved in viral entry into host cell

The GO terms are plotted in the UMAP, for retrieving protein functions for each GO ID, GO terms are shown in the table.