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

Dan Liu, Francesca Young, Kieran D. Lamb, David L. Robertson , Ke Yuan

LifeLU reading group

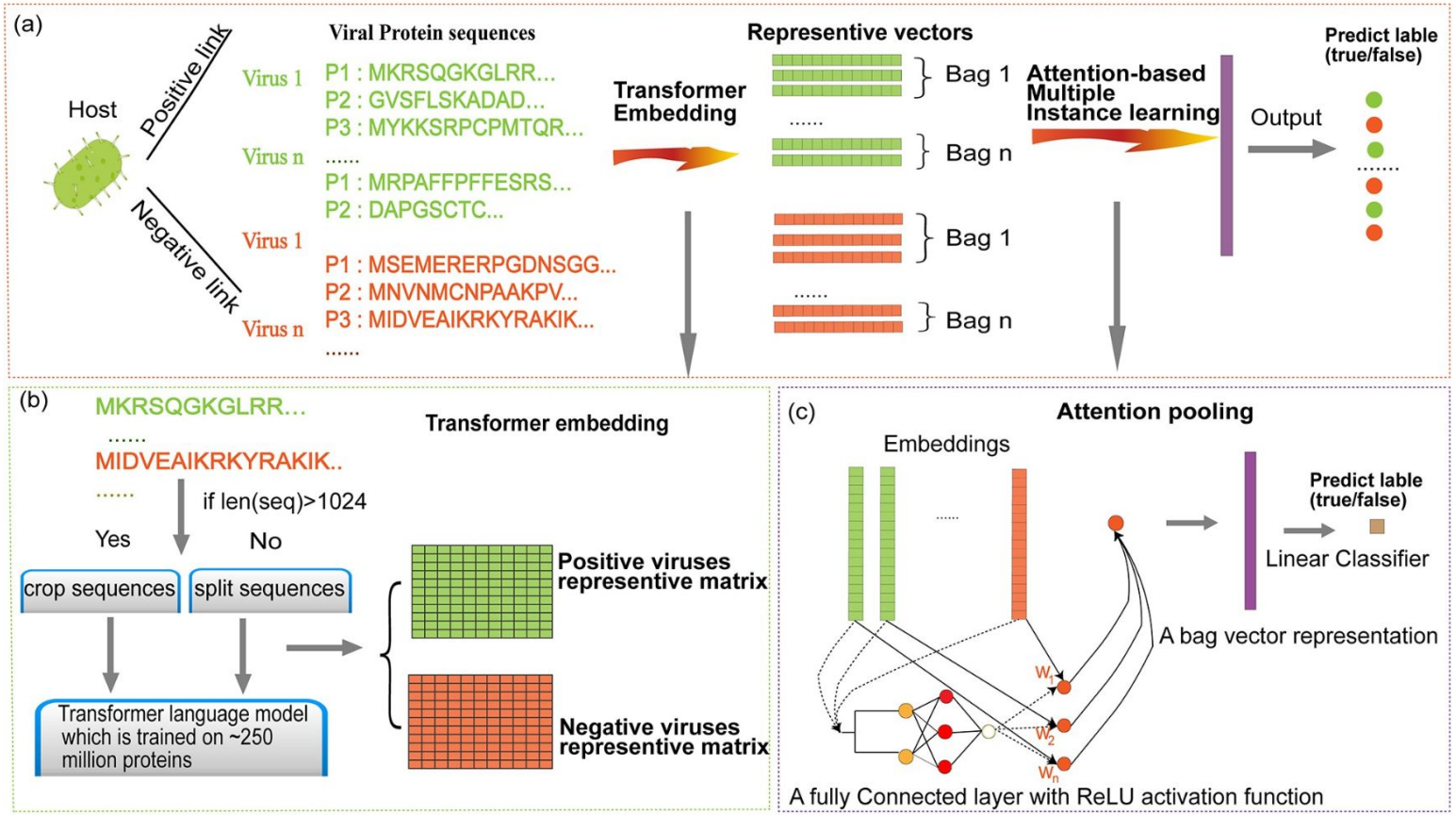
presented by Özdeniz Dolu

20.02.2025

The task: Virus-Host Association

- Viruses interact with some species (hosts) and don’t affect the others.
- %90 of virus sequences are not annotated with host information.
- The task is formulated both as a binary classification task for a particular host (interact or not) and a multi-class classification task for a set of hosts.

The architecture



Dataset and Data Preparation

- Data source: Virus-Host Database (VHDB)
- Two main datasets: *eukaryotes* and *prokaryotes*
- Data for 36 *eukaryote* hosts and 22 *prokaryote* hosts
- For each host, equal number of “positive” and “negative” samples.
- Each sample is a bag of protein sequences of a virus.
- How negative samples are selected depend on the *Strategy* (1 or 2).
- Only hosts with a number of known associations larger than a minimum threshold are selected.

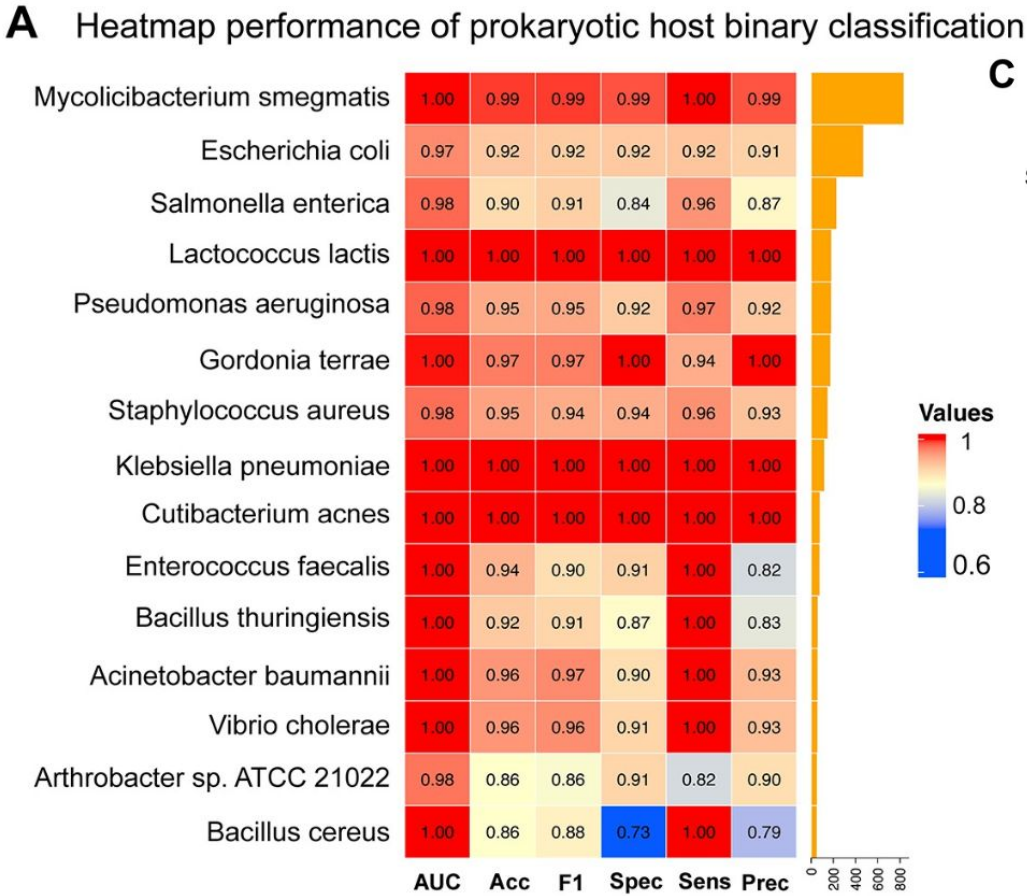
Dataset and Data Preparation

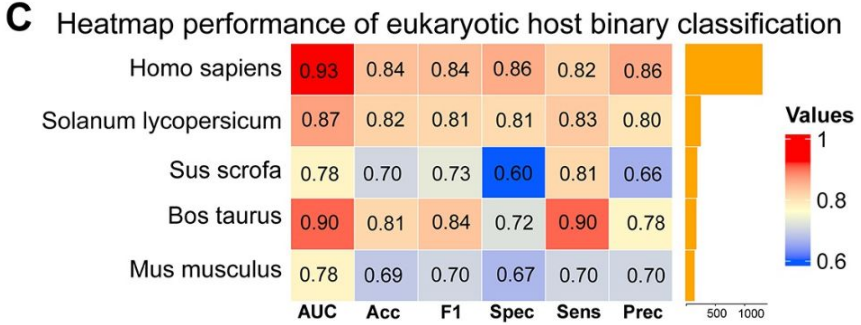
- Supplementary tables S1 and S2 show the statistics of the number of proteins and viruses associated with each host.

Binary Classification Experiment

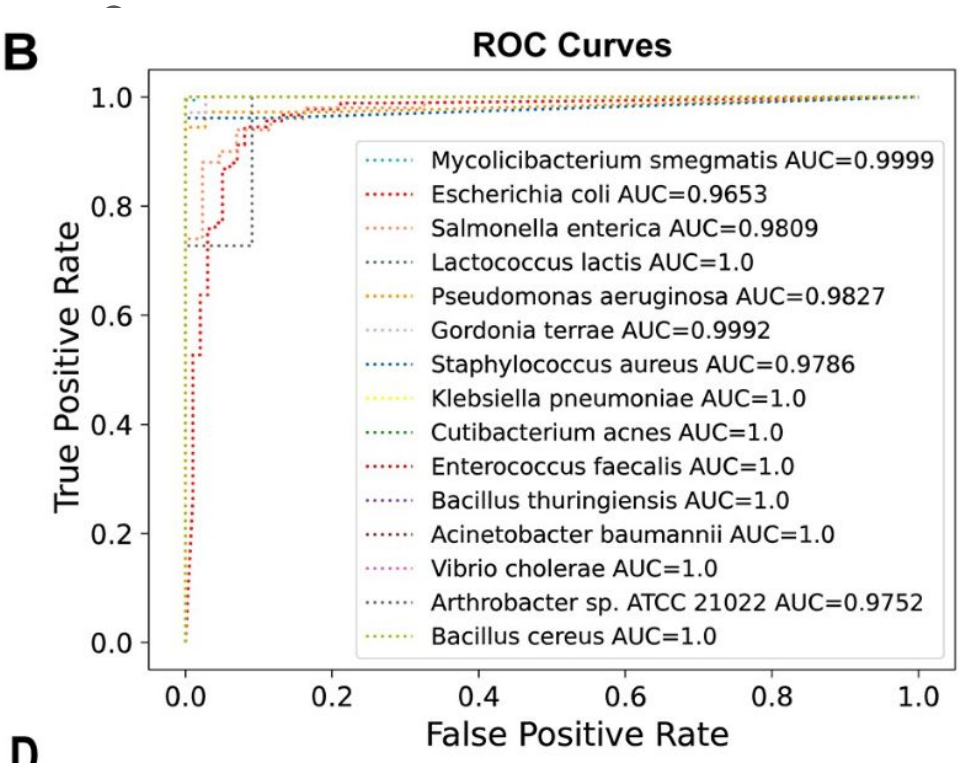
- Binary classifiers are trained on each host separately. (5-fold cross-validation).
- Negative samples for each host are selected with either Strategy 1 or Strategy 2.
- In Strategy 1, negative viruses are sampled from all viruses that are in different genera than positive viruses.
- In Strategy 2, negative viruses are sampled from viruses in same taxonomic class (for various classes) as the positive samples. Therefore, a harder task as the similarity increases.

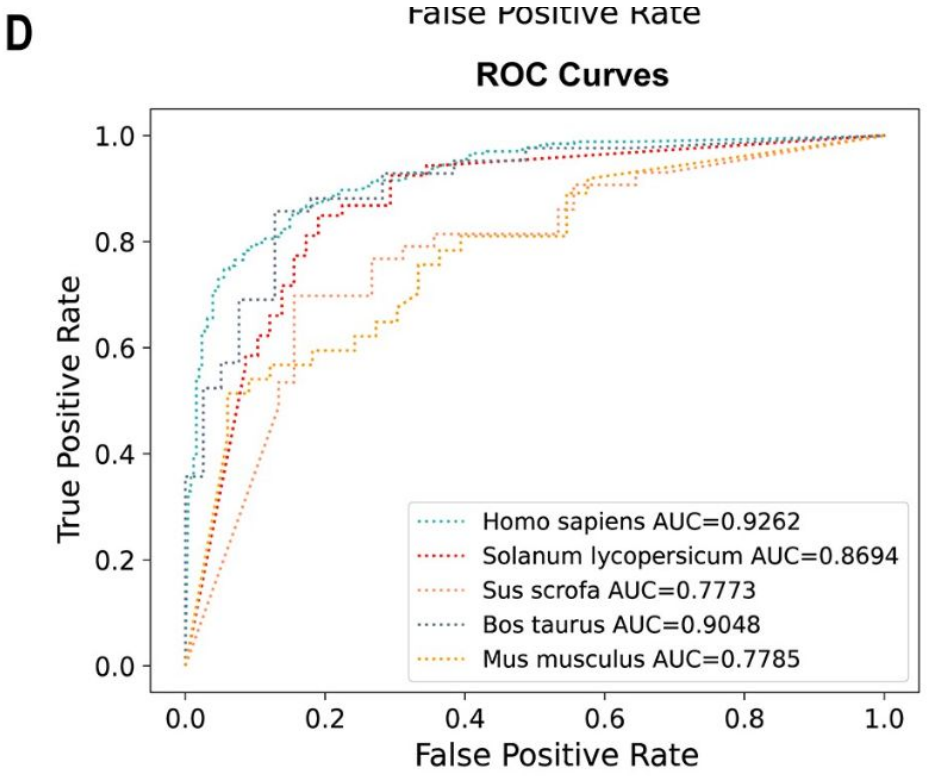
Binary Classification Experiment: Strategy 1





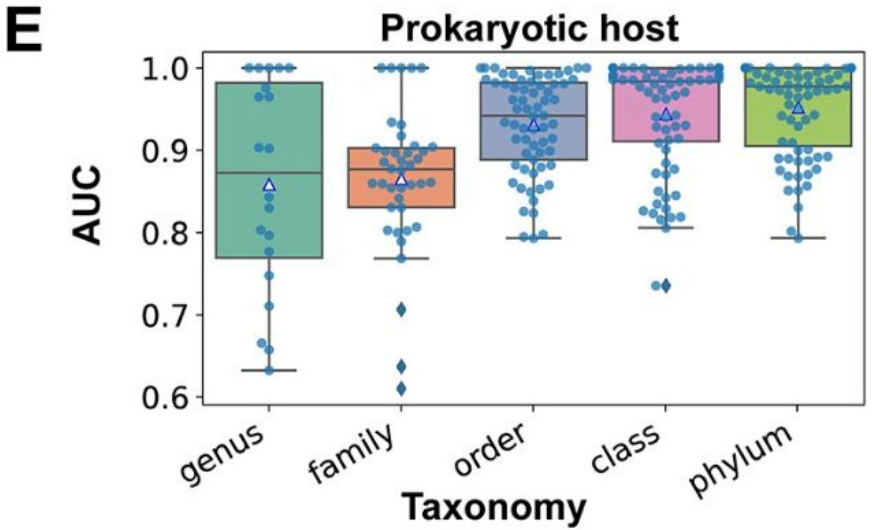
Binary Classification Experiment: Strategy 1

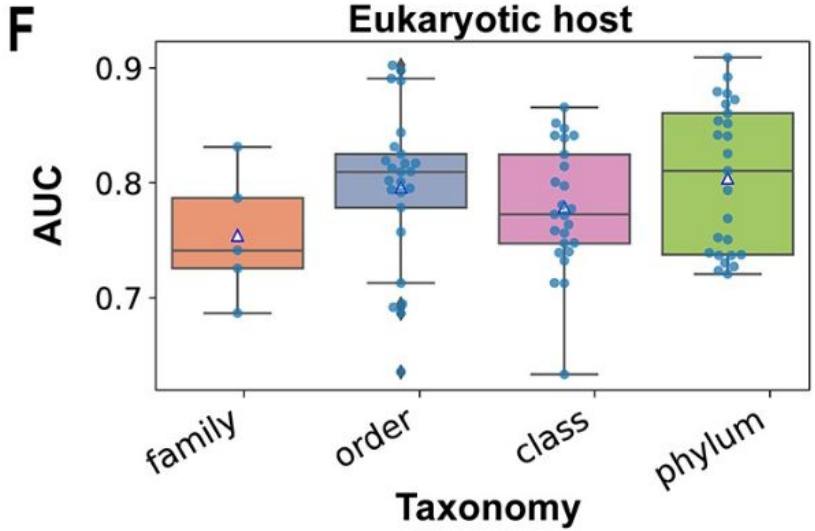




Binary Classification Experiment: Strategy 2

- When negative samples (viruses) are selected from the same genus, family, order, class or phylum respectively.





Binary Classification Experiment

- For eukaryotic host viruses, number of proteins are much smaller than that of prokaryotes. On top of that, virus types are much more varied for eukaryotic hosts. Therefore, it presents a harder classification challenge.
- In the case of Strategy 2, authors looked at the pairwise sequence similarity between positive and negative samples using MMSeq2 and they report that most of the scores are above 0.6.

Embedding vs k-mer Composition Experiment

- Two experiments for 22 prokaryotes and 36 eukaryotes.
- 5-fold cross validation.
- Multi-class classification task.
  - ESM-1b: embeddings are used as features.
  - AA_2: Amino-acid sequence k-mer composition is used as features.
  - PC_3: Physio-chemical sequence k-mer composition is used as features.
  - DNA_5: Nucleic acid sequence k-mer composition is used as features.

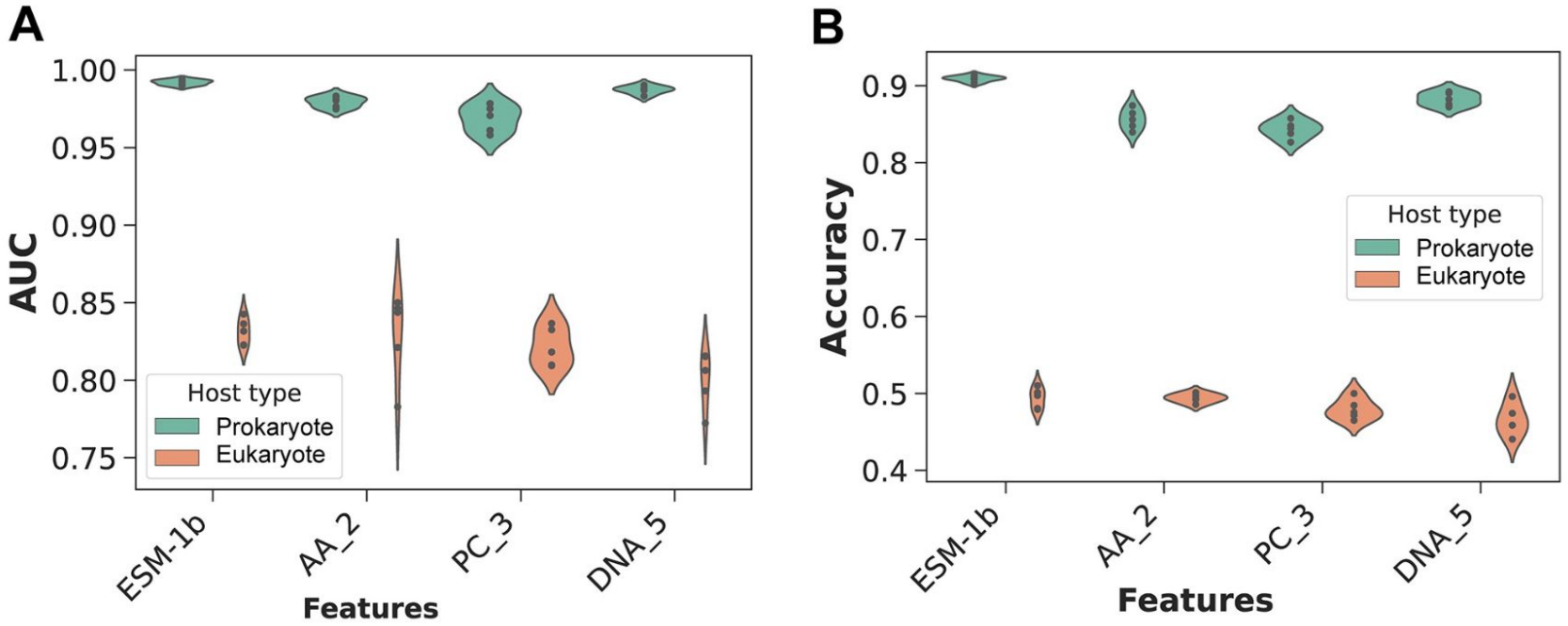
Embedding vs k-mer Composition Experiment

- ESM embeddings prove to be superior. However, DNA and AA based models are competitive.

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

Embedding vs k-mer Composition Experiment

- ESM embeddings prove to be superior. However, DNA and AA based models are competitive.

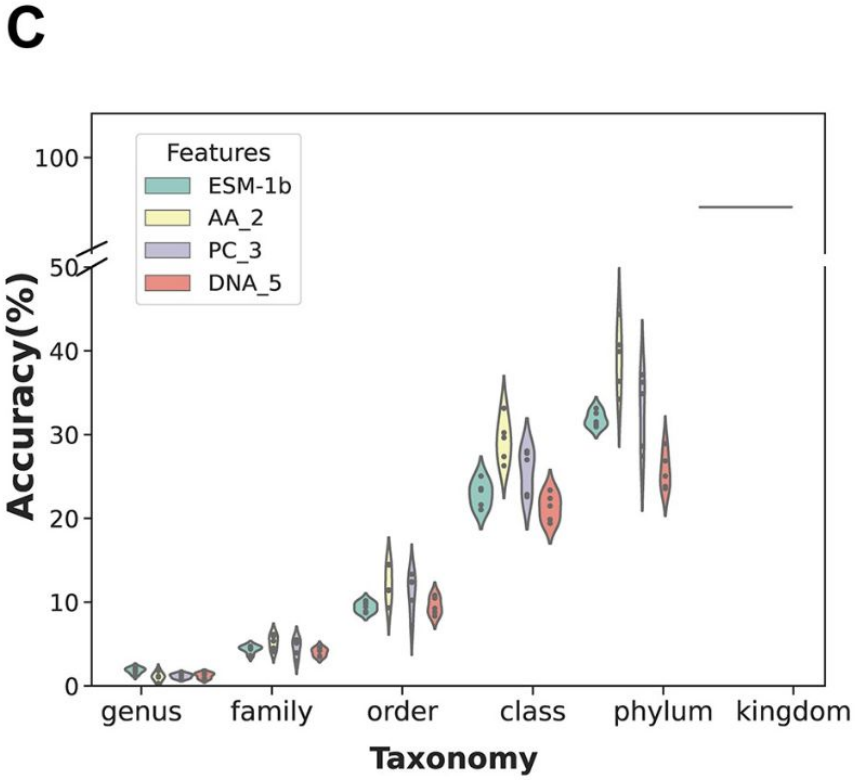


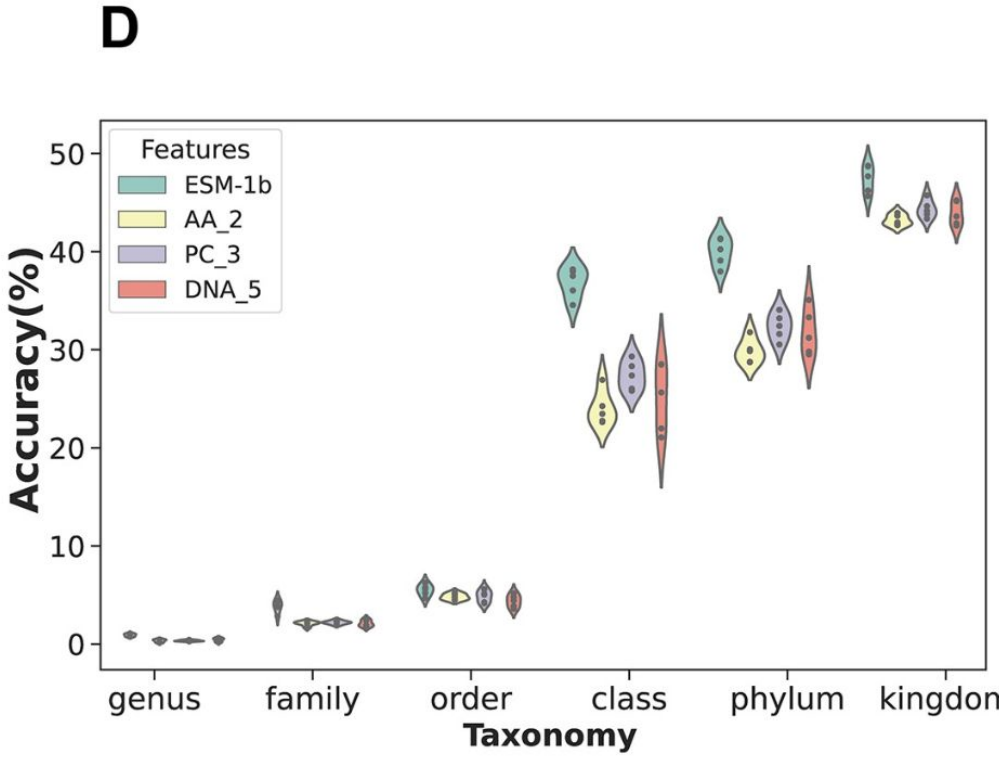
Embedding vs k-mer Composition Experiment 2

- For a harder challenge, viruses that have much less host annotations (5 to 30 hosts) whose annotations are not of any of the hosts in the training dataset are compiled from VHDB.
- Models have been evaluated on the resulting dataset.
- Various accuracy measures: A prediction is assumed “correct” if the predicted host is in same genus, family, order, class, phylum and kingdom.

Embedding vs k-mer Composition Experiment 2

- C: prokaryotes, D: eukaryotes, details are on S5 table





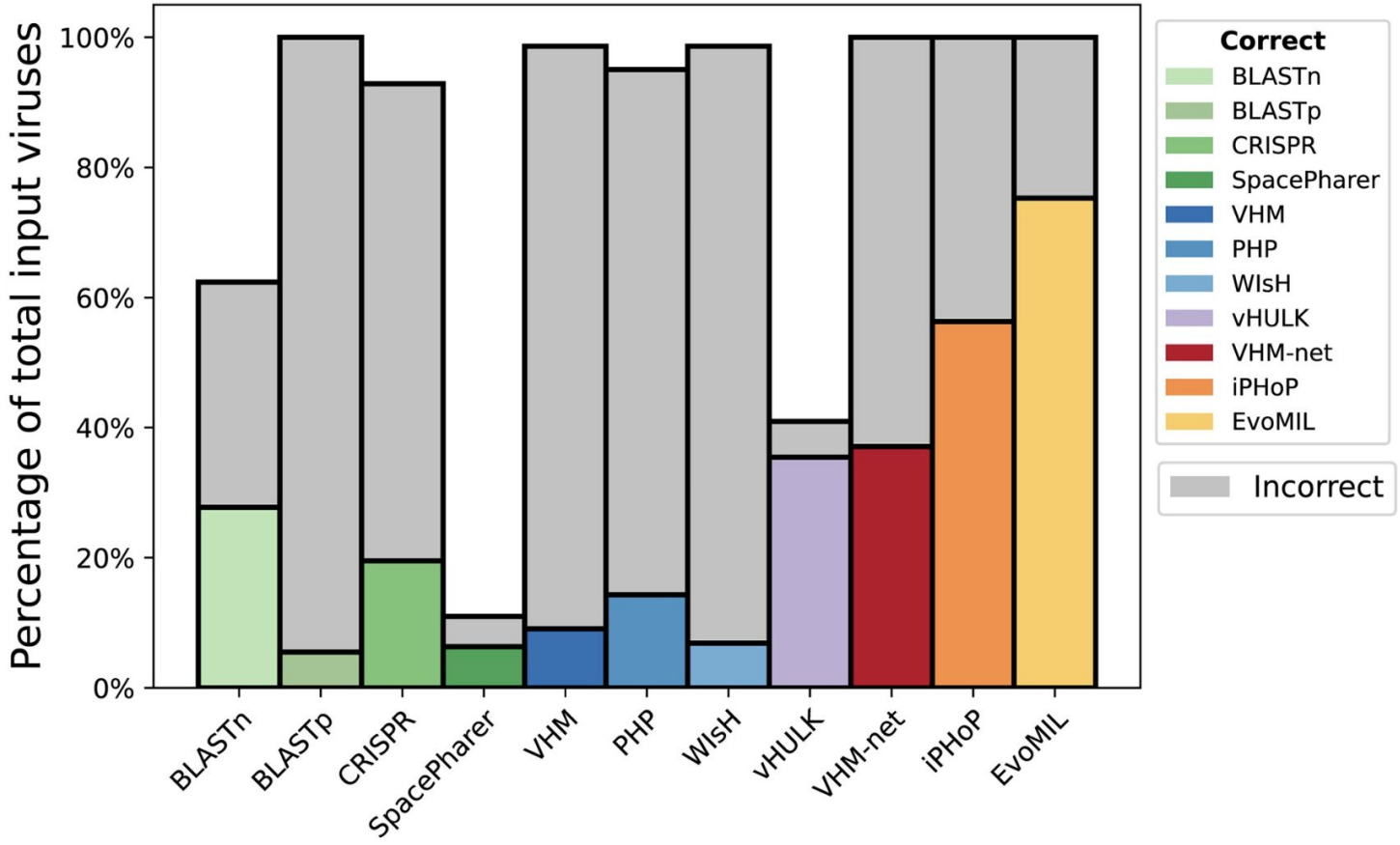
Embedding vs k-mer Composition Experiments

- Taxonomic tree aligned with accuracy ratios for each host, refer to the paper for detailed look (Fig 4, due to space constraints)
- It is also helpful to look at figure S4 for raw accuracy data (file s012.tif)
- S5 (s013.tif) contains a heatmap of “falsely predicted” classes for various taxonomy levels. (If the false prediction belongs to the same class, order, family etc.)
- Figure 5 contains confusion matrix for all classes (ESM-1b). (Helpful to analyze mistaken hosts)
- Main takeaway is that, mistakes of the models tend to be on the similar taxonomies.

Benchmarking vs Other Predictors

- 364 viruses are chosen across 22 prokaryotic hosts that EvoMIL is trained on. Compared with 9 prokaryotic host predictors iPHoP, BLASTn, BLASTp, CRISPR, WisH, VirHost-Matcher, PHP, SpacePHARER, VirHostMatcher-Net and vHULK. (Full list on Table S7). Data is taken from test dataset of iPHoP paper.

Benchmarking vs Other Predictors



MIL Attention Weights Analysis

- Each protein of a virus is associated with a weight. A higher weight suggest a higher participation in the prediction.
- For every virus associated with E. Coli and H. Sapiens, top 5 ranked (in terms of weight) proteins have been collected and globally ranked. Also, weight ranking of **all proteins** have also been included.
- GO annotations of these proteins have been analyzed.
- Figures show results for both binary classifier model and multi class classifier model.

MIL Attention Weights Analysis (Refer to Figure 7)

- In Figure 7, we see that compiling the top 5 ranking proteins captures the important GO annotations because we don’t miss these GO annotations when ignoring lesser ranking proteins. (???)
- E. Coli and H. Sapiens are selected because they are extensively studied and have the most annotations.

MIL Attention Weights Analysis (Refer to Figure S6)

- A dendrogram (wrt. hierarchical clustering of ESM embeddings) of top-ranking proteins with viral life cycle GO annotations is provided in this figure.
- There’s an apparent clustering with respect to various GO terms, implying that ESM embeddings encode functional information although there are some exceptions.

MIL Attention Weights Analysis (SARS-CoV-2 case study)

- The virus SARS-CoV-2 is unseen by the EvoMIL. H. Sapiens binary classifier is investigated on this sample.
- EvoMIL predicted its host as H. Sapiens with 0.97 probability whereas k-mer SVM classifier gave 0.01.
- Top three ranking proteins (wrt. attention weights) were spike subsequences, non-structural proteins subsequences and Nucleocapsid protein.
- All three have been associated with viral process GO terms and play a role in interaction with the host.
- This case is provided as evidence for capability of EvoMIL in identifying important proteins in virus-host associations.

Thanks for listening