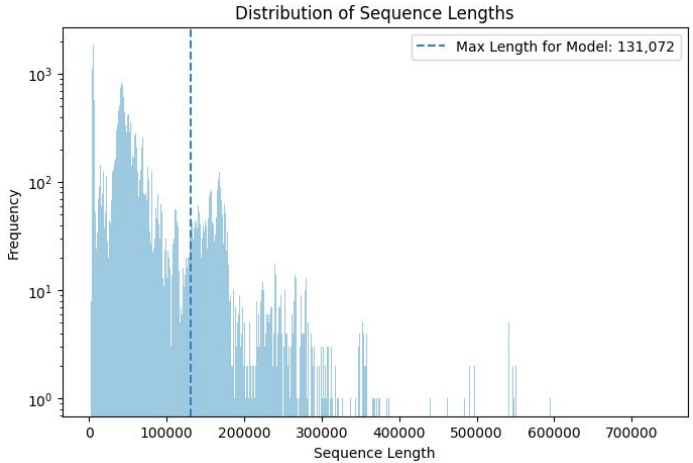
LLMs for phages, deliverable 1

by Xinlu Shi

Dataset

- Dataset: sequences of phage genomes and the metadata from https://millardlab.org/phage-genome-jan2025/
- Number of sequences in the dataset: 32,043
- Sequence length of genomes in the dataset:

count	32,043
mean	60,216
std	55,525
min	1,761
25%	33,533
50%	44,866
75%	67,558
max	735,411



* When training the foundation model ("megaDNA"), they used a filtered dataset consisting of sequences < 96k bp;
* The max sequence length the model expects as input is 131,072 (2^17) bp.

Embeddings - dataset filtering

The objective is to assess whether embeddings align with metadata, so I filtered the dataset based on the following criteria:

- Retained only phages with complete metadata for host, genus, and family.
- Included only phages with a sequence length within the model's expected maximum length (131,072).
- After filtering, 8,069 phages remain.

Accession	Genus	Family	Host	Genome Length (bp)
AY319521	Felsduovirus	Peduoviridae	Salmonella	35155
AC171169	Tequintavirus	Demerecviridae	Escherichia	104373
AY576273	Keylargovirus	Mesyanzhinovviridae	Alphaproteobacteria	63649
MN335248	Xylivirus	Inoviridae	Vibrio	7045
MG592615	Livvievirus	Autolykiviridae	Vibrio	10611

Identifiers for phages

Embeddings - constructing feature vectors

- I followed the same approach as described in the paper:

  “Model embeddings were extracted from three layers (dim = 196, 256, 512) and concatenated together to form a 964-dim vector for each input sequence.”
  https://www.nature.com/articles/s41467-024-53759-4

Comparison on clustering result and metadata

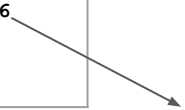
- Perform clustering based on embeddings.
- Experiment with different params for clustering: PCA dimensions, clustering methods, and the number of clusters.
- For each cluster, compute purity scores (percentages) based on the most frequent genus, family, or host within the cluster.
- Higher purity scores indicate stronger embeddings-metadata association.

Comparison on clustering result and metadata - Results

* Number of different **Families/Genera/Hosts** in the filtered dataset: **73/512/133**

Some of the results with different clustering configurations:

Clustering Method	Hierarchical (sklearn.cluster.AgglomerativeClustering)	Hierarchical (sklearn.cluster.AgglomerativeClustering)	Hierarchical (sklearn.cluster.AgglomerativeClustering)
number of clusters	500	100	10
PCA dimensions	964 (original dim)	964(original dim)	100
Results	explained variance: 1.000 Genus Purity: **0.830** Family Purity: **0.985** Host Purity: **0.821**	explained variance: 1.000 Genus Purity: **0.669** Family Purity: **0.927** Host Purity: **0.677**	explained variance: 0.966 Genus Purity: **0.397** Family Purity: **0.656** Host Purity: **0.368**



This can be considered high purity score given that there are 73 families in total yet the number of cluster is only 10.

Comparison on clustering result and metadata - Results

* Number of different **Families/Genera/Hosts** in the filtered dataset: **73/512/133**

Some samples of clusters:

```
================================================
Clustering Method: Hierarchical (n_clusters=500)
PCA dimensions: 964 (explained variance: 1.000)
================================================

Cluster 0 (Size: 23)

Top Genera:
Genus
Microvirus    23

Top Families:
Family
Microviridae    23

Top Hosts:
Host
bajarodmic    15
Dulem          8
```

```
Cluster 65 (Size: 69)

Top Genera:
Genus
Epseptimavirus    54
Tequintavirus     15

Top Families:
Family
Demerecviridae    69

Top Hosts:
Host
Salmonella     60
Escherichia     9
```

```
================================================
Clustering Method: Hierarchical (n_clusters=100)
PCA dimensions: 964 (explained variance: 1.000)
================================================

Cluster 5 (Size: 100)

Top Genera:
Genus
Teseptimavirus    100

Top Families:
Family
Autographiviridae    100

Top Hosts:
Host
Escherichia    96
Yersinia        2
Cedecea         1
Enterobacteria  1
```

```
Cluster 11 (Size: 111)

Top Genera:
Genus
Microvirus              101
Inovirus                  3
Affertcholeramvirus       2
Infulavirus               2
Emesvirus                 1

Top Families:
Family
Microviridae      101
Inoviridae          7
Fiersviridae        1
Demerecviridae      1
Cystoviridae        1

Top Hosts:
Host
bajarodmic    101
Escherichia     4
Dulem           3
Vibrio          2
Pseudomonas     1
```

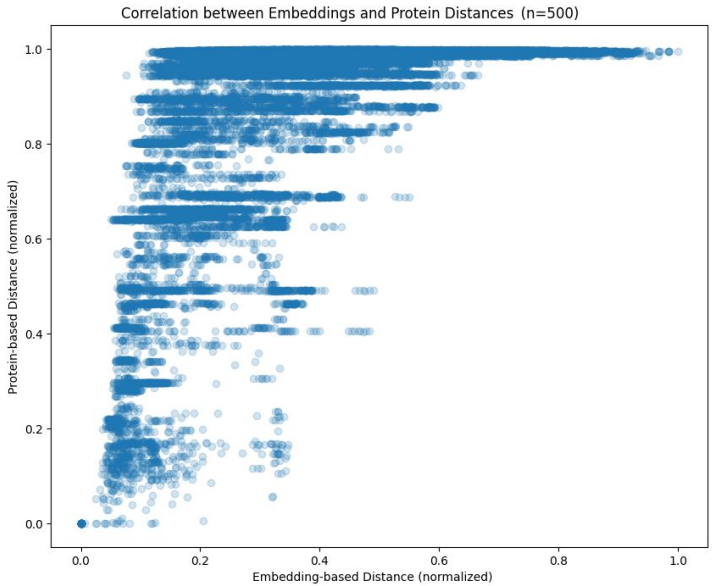
```
Cluster 16 (Size: 128)

Top Genera:
Genus
Microvirus        122
Tertilicivirus      4
Vicialiavirus       2

Top Families:
Family
Microviridae    122
Inoviridae        6

Top Hosts:
Host
bajarodmic    86
Dulem         36
Pseudomonas    4
Vibrio         2
```

Conclusion: embeddings are informative in the context of both taxonomy and viral host, esp. in taxonomy

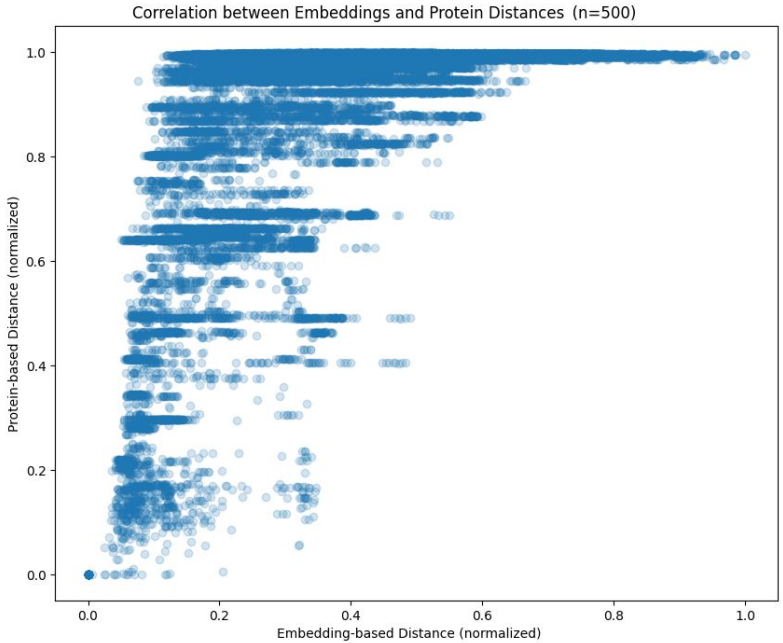
Comparison on Embeddings & Protein Similarity

- **Embeddings-Based Distance:** Trying both Euclidean distance and cosine similarity.
- **Protein-Based Distance:** Calculated as the sum of branch lengths between two nodes in the tree (tree proteomic distance).
- **Normalization:** Scaling distances to the range (0,1).

- `number of common accessions between the 2 datasets: 1375`
- `500 phages is sampled`
- `combination(500,2) = 124,750 pairs of phages`



Comparison on Embeddings & Protein Similarity



Observation:

- **No obvious linear correlation**
- **similar** protein (normalized distance < 0.2) → **similar** embeddings (normalized distance < 0.4)
- **different** embeddings (normalized distance > 0.6) → **really different** protein (normalized distance close to 1)
- **similar** embeddings (normalized distance < 0.2) → protein distance may **vary** (normalized distance between (0,1))