Deliverable 1. Modeling on embeddings generated through megaDNA_145M and evo2_1b_base

Overview

I generated embeddings using the foundation models MegaDNA and Evo2 and performed clustering to assess whether the embeddings grouped according to labels such as phage or host taxonomy. Additionally, I compared the distances between embeddings with tree proteomic distances for pairs of phage genome sequences. To further evaluate the embeddings, I trained MLP models to predict viral taxonomy and host information. These models were trained on the Millardlab dataset (using 90% of the data for training and 10% for validation) and tested on the MATRIX phages dataset.

Table: Comparison between megaDNA_145M and evo2_1b_base

Model		megaDNA_145M	evo2_1b_base
Accuracy in **Class**	train	99.90%	89.40%
	val	100%	90.43%
	test	79.36%	33.83%
Accuracy in **Family**	train	99.42%	88.24%
	val	99.75%	89.81%
	test	65.06%	14.29%
Accuracy in **Genus**	train	93.48%	81.42%
	val	94.50%	81.79%
Accuracy in **Host**	train	86.15%	68.18%
	val	86.75%	67.59%
	test	9.16%	1.12%

Dataset

Millardlab dataset

https://millardlab.org/2025/03/06/phage-genomes-march-2025/

Sequence lengths:

```
count     32043.000000 #32,043 genomes in total
mean      60215.966607
std       55524.968903
min        1761.000000
25%       33532.500000
```

```
50%       44866.000000
75%       67557.500000 #lengths mostly below 100k bp
max      735411.000000
```

Metadata available:

'Accession', 'Description', 'Classification', 'Genome Length (bp)',
'Jumbophage', 'molGC (%)', 'Molecule', 'Modification Date',
'Number CDS', 'Positive Strand (%)', 'Negative Strand (%)',
'Coding Capacity (%)', 'Low Coding Capacity Warning', 'tRNAs', **'Host'**,
'Lowest Taxa', **'Genus'**, **'Sub-family'**, **'Family'**, **'Order'**, **'Class'**,
'Phylum', **'Kingdom'**, **'Realm'**, 'Baltimore Group', 'Genbank Division',
'Isolation Host (beware inconsistent and nonsense values)'

- The entries in the column “Host” are mostly `Genera` of viral hosts, but can sometimes be `Phyla/Classes/Families.`

Field	Percentage of information being present (not “Unspecified” or “Unclassified”)
Host	90.48%
Genus	71.14%
Family	44.68%
Order	18.40%
Class	96.07%
Phylum	96.07%

MATRIX dataset

Sequence lengths:

```
count      9345.000000 # 9,345 genomes in total
mean      24201.212841
std       19626.124086
min        2196.000000
25%       12655.000000
50%       17409.000000
75%       29948.000000 # mostly below 30k bp
max      354857.000000
```

Metadata available:

'vOTU', 'isolate', 'microbial_fraction', 'length', 'A', 'C', 'G', 'T',
'GC', 'geNomad_viral_conservative', 'geNomad_genetic_code',
'geNomad_virus_score', 'geNomad_n_hallmarks', **'geNomad_taxonomy'**,
'geNomad_topology', 'geNomad_n_genes', 'geNomad_provirus',
'VIBRANT_lifecycle', 'VirSorter2_max_score', 'checkv_quality',
'checkV_completeness', 'checkV_completeness_method',
'checkV_contamination', 'checkV_kmer_freq', 'checkV_warnings',
'blast_sseqid', 'blast_salltitles', 'blast_qcovs', 'blast_length',
'PhaGCN_prediction', 'PhaGCN_score', 'iphop_confidence_score',
'iphop_phylum', **'iphop_class'**, **'iphop_order'**, **'iphop_family'**,
'iphop_genus', 'DENMARK', 'LEAF', 'DENMARK_LEAF',

'DENMARK_PHYLLOSPHERE', 'USA_LEAF', 'RIZOSPHERE_DENMARK', 'WHEAT_RIZOSPHERE_DENMARK'

- We have genus/family/order/class/phylum for host
- For viral taxonomy, “Class” information are mostly present; 98% are with the same Class “Caudoviricetes”. “Order”, “Family” and “Genus” information are mostly absent.

Embeddings: megaDNA

- Get embeddings: megaDNA_145M; Millardlab dataset
- Concatenate the embeddings from the three layers into one 964-dim vector

Clustering

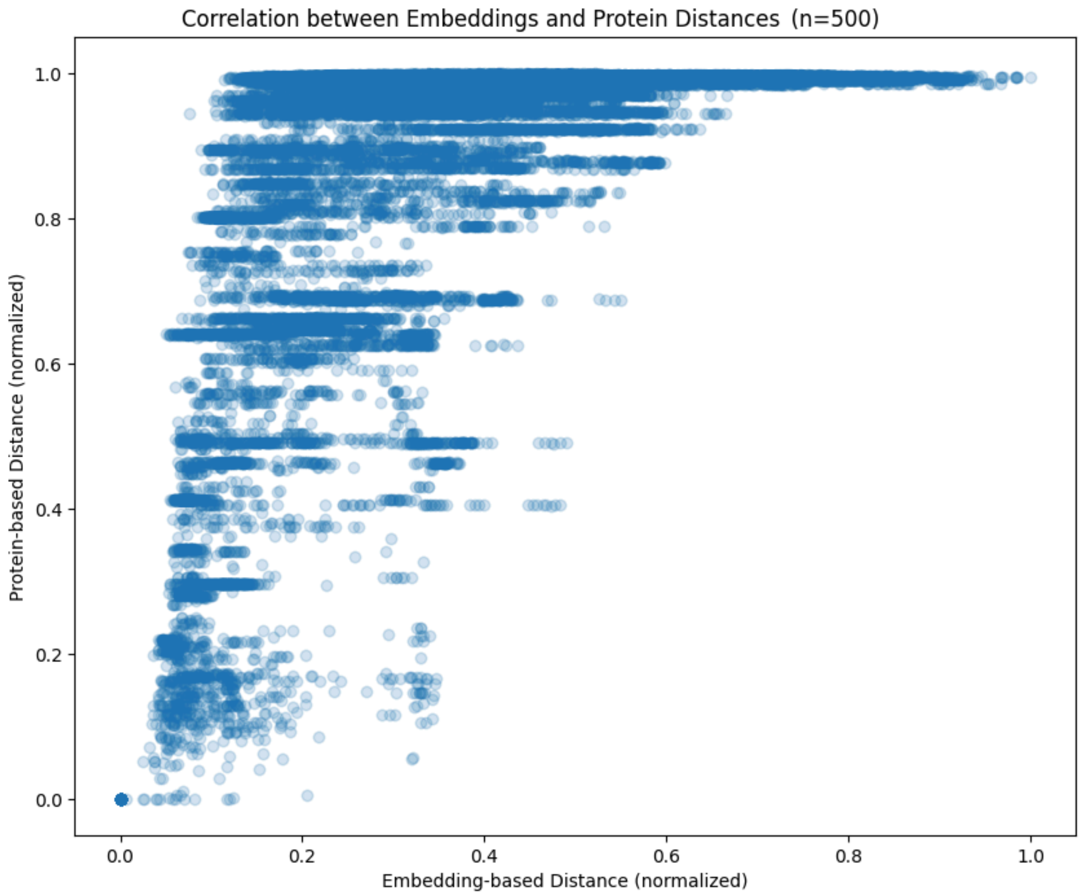
Observation: The embeddings are closely related to taxonomy and viral hosts, esp. family.

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**

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

Distance in embeddings vs. proteins

- 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).



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))

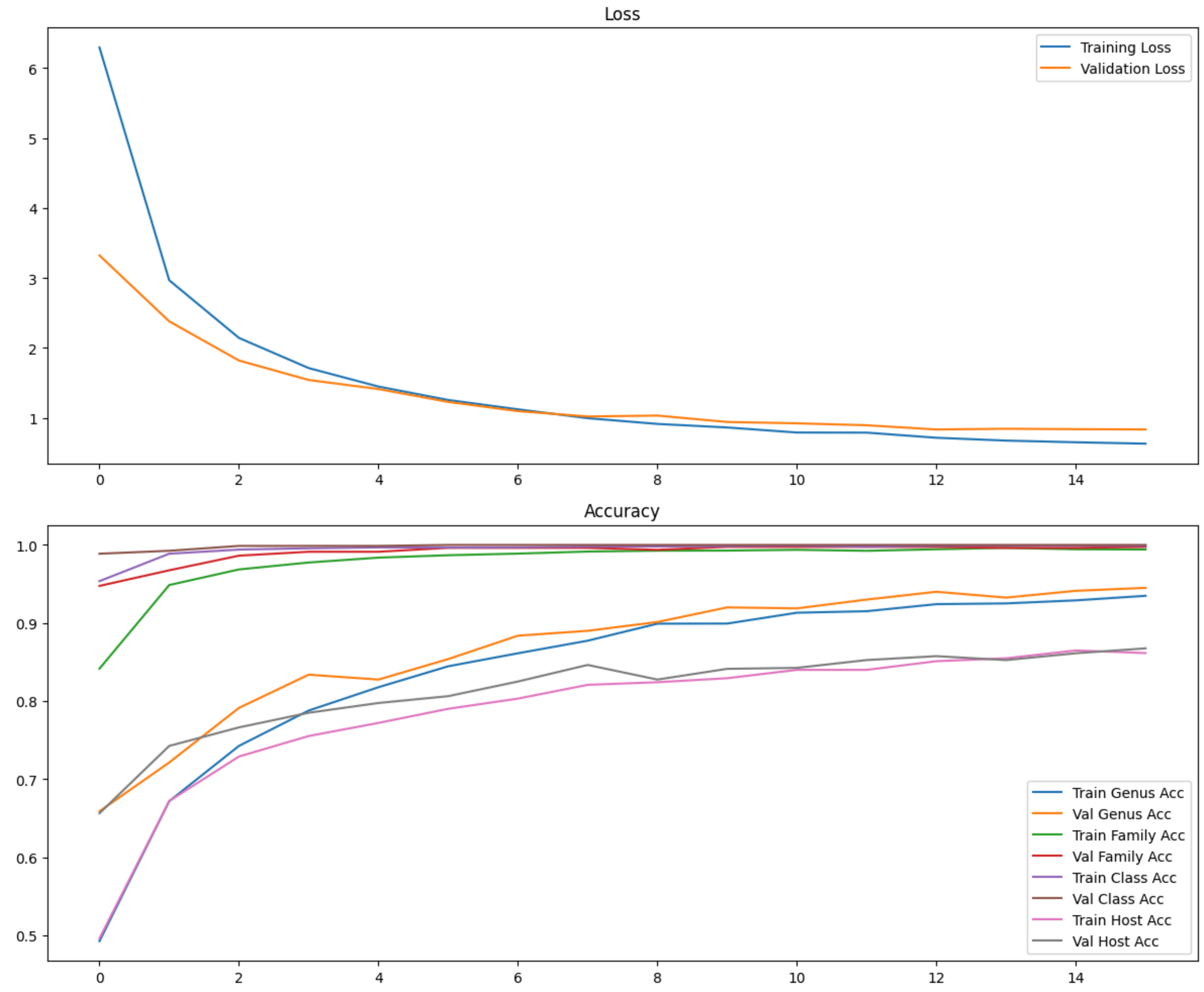
MLP model

Implementation:

- Takes the embeddings as input and predict viral Class/Family/Genus/Host. (For host, only genus of host is predicted due to limitation of training data)
- Trained on Millardlab dataset (90% for training and 10% for validation)
- Tested on the MATRIX dataset. (only Class/Family/Host predictions are tested because MATRIX metadata lacks genus information)

Observation:

- Training/Validation: accuracy close to 100% for Class and Family prediction; accuracy close to 0.9 for Genus prediction; accuracy close to 0.8 for Host prediction
- Testing: the dataset for testing is imbalanced; accuracy is considerably good for predicting Class (79%); accuracy becomes worse when predicting Family (65%), possibly due to too few samples available; accuracy is only 9.16% when predicting Host.



```
--- Testing on CLASS prediction ---
Number of test samples: 9210
Class prediction accuracy: 0.7936
                  precision    recall  f1-score   support

  Caudoviricetes       1.00      0.79      0.88      9201
  Faserviricetes       0.01      1.00      0.01         4
Tectiliviricetes       0.01      1.00      0.03         5

       micro avg       0.88      0.79      0.83      9210
       macro avg       0.34      0.93      0.31      9210
    weighted avg       1.00      0.79      0.88      9210


--- Testing on FAMILY prediction ---
Number of test samples: 135
* 52 samples have unknown labels (not seen during training) and will be excluded
Family prediction accuracy: 0.6506
                   precision    recall  f1-score   support

Autographiviridae       0.90      0.72      0.80        61
   Drexlerviridae       0.67      1.00      0.80         2
   Herelleviridae       0.00      0.00      0.00         4
       Inoviridae       1.00      1.00      1.00         4
     Tectiviridae       0.75      0.60      0.67         5
    Zobellviridae       1.00      0.14      0.25         7

        micro avg       0.89      0.65      0.75        83
        macro avg       0.72      0.58      0.59        83
     weighted avg       0.85      0.65      0.72        83


--- Testing on HOST prediction ---
Number of test samples: 9305
* 7066 samples have unknown labels (not seen during training) and will be excluded
Host prediction accuracy: 0.0916
                  precision    recall  f1-score   support

   Achromobacter       0.00      0.00      0.00         7
   Acinetobacter       0.00      0.00      0.00         4
   Agrobacterium       0.00      0.00      0.00        44
        Bacillus       0.09      1.00      0.17         1
     Bacteroides       0.00      0.00      0.00        21
      Bordetella       0.00      0.00      0.00         1
     Citrobacter       0.00      0.00      0.00        10
     Clavibacter       0.00      0.00      0.00        13
 Corynebacterium       0.00      0.00      0.00         1
     Cronobacter       0.00      0.00      0.00         6
    Enterobacter       0.00      0.00      0.00        24
    Enterococcus       1.00      0.12      0.22         8
         Erwinia       0.22      0.05      0.08       106
     Escherichia       0.02      0.14      0.03        22
 Exiguobacterium       0.00      0.00      0.00         2
  Flavobacterium       1.00      0.62      0.76        13
        Gordonia       0.08      0.33      0.12         3
          Hafnia       0.00      0.00      0.00        50
      Klebsiella       0.03      0.32      0.06        25
       Kosakonia       0.00      0.00      0.00         1
   Lactobacillus       0.00      0.00      0.00         1
       Leclercia       0.00      0.00      0.00         5
   Mesorhizobium       0.00      0.00      0.00         4
  Microbacterium       0.29      0.18      0.22        28
      Morganella       0.00      0.00      0.00         2
   Mycobacterium       0.21      0.02      0.04       131
          Nostoc       0.00      0.00      0.00         5
         Pantoea       0.00      0.00      0.00       260
  Pectobacterium       0.00      0.00      0.00         2
         Proteus       0.00      0.00      0.00         3
     Providencia       0.00      0.00      0.00         8
     Pseudomonas       0.63      0.29      0.40       572
       Rhizobium       0.00      0.00      0.00        34
     Rhodococcus       0.00      0.00      0.00        45
      Salmonella       0.00      0.00      0.00        11
        Serratia       0.00      0.00      0.00        20
   Sinorhizobium       0.00      0.00      0.00         6
    Sphingomonas       0.00      0.00      0.00       664
Stenotrophomonas       0.01      0.18      0.03        11
    Streptomyces       0.00      0.00      0.00        41
          Vibrio       0.02      0.11      0.04         9
     Xanthomonas       0.25      0.50      0.33         2
        Yersinia       0.02      0.08      0.03        13

       micro avg       0.18      0.09      0.12      2239
       macro avg       0.09      0.09      0.06      2239
    weighted avg        0.20       0.09       0.12       2239
```

Embeddings: evo2

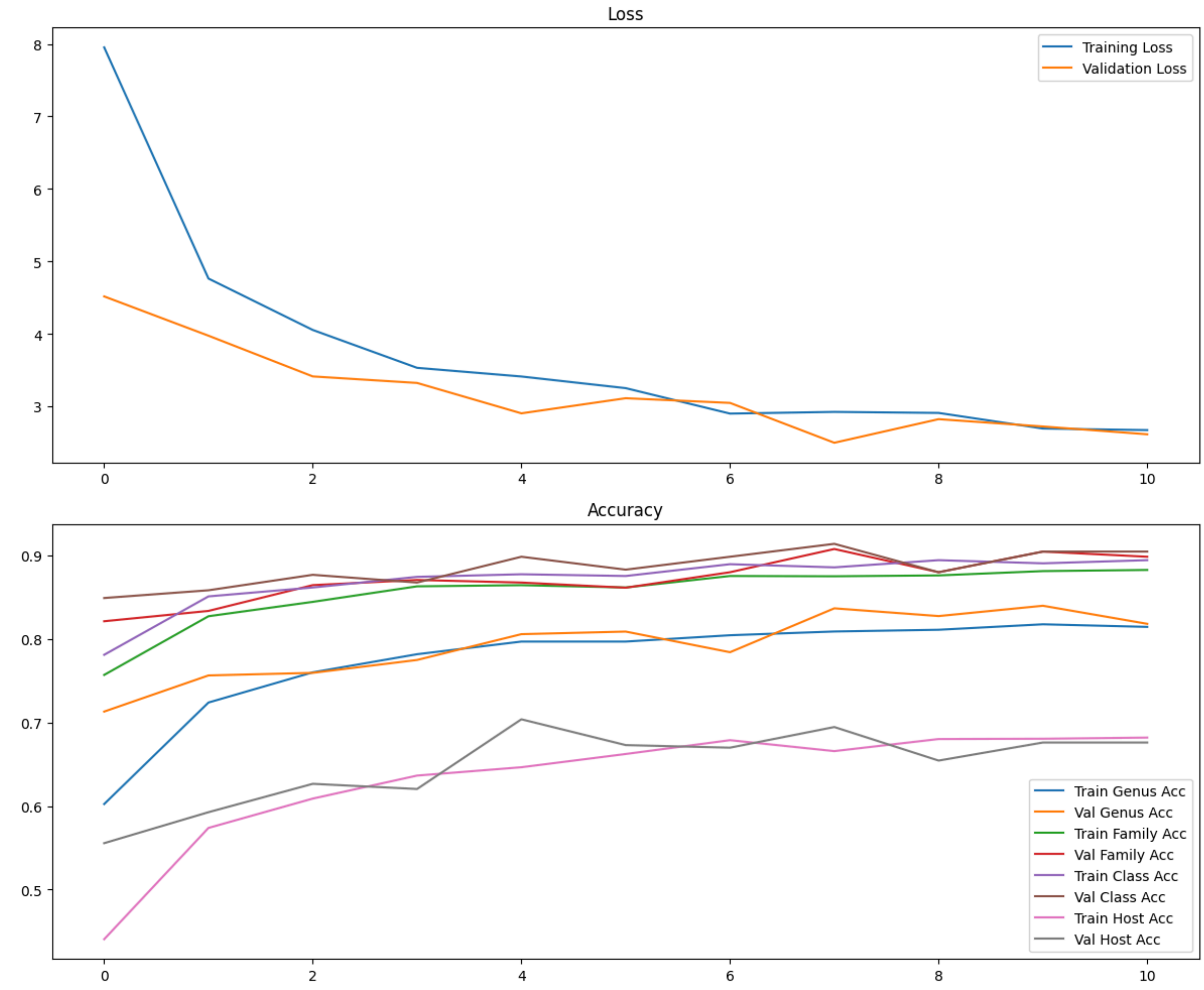
Implementation:

- The medium version `(evo2_7b)` needs H100 and the largest version `(evo2_40b)` needs 2 x H100. For now I just present the results I got with the smallest version `(evo2_1b_base)` and short sequences (<30k bp) from the dataset due to computational limitations. The smallest version is pretrained with 8192 context length so can have limited performance.
- The shape of original embeddings is [1,n,1920] where n is the length of the sequence. I averaged over the sequence length dimension to get a 1920-dim feature vector for each sequence.

MLP model

Observation:

- Training/Validation: even the smallest version of evo2 has considerably high validation accuracy: 90% for Class prediction, 88% for Family prediction, 80% for Genus prediction, 65% for Host prediction.
- Testing: the testing accuracy for evo2_1b_base is considerably worse than megaDNA.



```
--- Testing on CLASS task ---
Number of test samples: 1008
Class prediction accuracy: 0.3383
```

```
                precision    recall  f1-score   support

 Caudoviricetes       1.00      0.34      0.51      1004
 Faserviricetes       0.00      0.25      0.01         4

      micro avg       0.61      0.34      0.44      1008
      macro avg       0.50      0.29      0.26      1008
   weighted avg       1.00      0.34      0.50      1008


--- Testing on FAMILY task ---
Number of test samples: 16
*9 samples have unknown labels and will be excluded
Family prediction accuracy: 0.1429
                   precision    recall  f1-score   support

Autographiviridae       0.00      0.00      0.00         2
   Herelleviridae       0.00      0.00      0.00         1
       Inoviridae       0.33      0.25      0.29         4

        micro avg       0.33      0.14      0.20         7
        macro avg       0.11      0.08      0.10         7
     weighted avg       0.19      0.14      0.16         7


--- Testing on HOST task ---
Number of test samples: 1024
*935 samples have unknown labels and will be excluded
Host prediction accuracy: 0.0112
                  precision    recall  f1-score   support

     Cronobacter       0.00      0.00      0.00         1
         Erwinia       0.00      0.00      0.00        10
     Escherichia       0.00      0.00      0.00         2
      Klebsiella       0.00      0.00      0.00         2
  Microbacterium       0.05      0.50      0.09         2
     Pseudomonas       0.00      0.00      0.00        57
     Rhodococcus       0.00      0.00      0.00         6
Stenotrophomonas       0.00      0.00      0.00         3
    Streptomyces       0.00      0.00      0.00         3
          Vibrio       0.00      0.00      0.00         2
     Xanthomonas       0.00      0.00      0.00         1

       micro avg       0.04      0.01      0.02        89
       macro avg       0.00      0.05      0.01        89
    weighted avg       0.00      0.01      0.00        89
```