Deliverable 2. MLP with different training strategies

Overview

Features:

1. use new datasets with metadata containing both viral and host taxonomy.
2. use the largest version of megaDNA (megaDNA_277M)
3. use manually implemented hierarchical loss, which simply gives more penalty when the predictions for higher level taxonomy are wrong
4. train the MLP model with 2 different strategies: 1) predict for the whole test set; 2) split the test set in two halves and give one half to the training. I did this for both MATRIX and STRAIN dataset.

Observation: The model predicts viral/host taxonomy much better with the dataset-splitting strategy. The result on host taxonomy prediction is especially good on STRAIN where ground-truth hosts are known.

Dataset

Millardlab dataset (ver.Mar)

Sequence lengths:

```
count     33166.000000 # 33,166 genomes in total
mean      60853.816348
std       55336.412598
min        1761.000000
25%       34593.000000
50%       45404.000000
75%       68389.000000
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)',
```

```
'Host_domain', 'Host_phylum', 'Host_class', 'Host_order', 'Host_family',
'Host_genus'
```

Percentage of available viral/host taxonomy. For viral taxonomy, we can sometimes know the lower level yet not know the higher level, e.g. there are more phages remaining unclassified on Family than Genus.

```
Genus Unclassified: 0.31174430875923415
Sub-family Unclassified: 0.6947384290667873
Family Unclassified: 0.5673149404492688
Order Unclassified: 0.8221317654153475
Class Unclassified: 0.06941052314186642
Phylum Unclassified: 0.06941052314186642
Kingdom Unclassified: 0.06941052314186642
Realm Unclassified: 0.06941052314186642

Host_genus is nan: 0.12069953263983114
Host_family is nan: 0.12045831448816524
Host_order is nan: 0.12015679179858284
Host_class is nan: 0.1201266395296246
Host_phylum is nan: 0.12003618272274989
Host_domain is nan: 1.0
```

MATRIX dataset (ver. Mar)

viral sequences assembled from metagenomes; sequences are from viruses found in the phyllosphere; ground-truth host not known

Sequence lengths:

```
count      8399.000000 #8,399 genomes in total
mean      24608.987499
std       20124.994540
min        4441.000000
25%       12782.000000
50%       17745.000000
75%       30802.500000
max      354857.000000
```

Metadata available:

```
'vOTU', 'length', 'geNomad_viral_conservative', 'checkv_quality',
'iphop_host_confidence_score',
'iphop_host_phylum', 'iphop_host_class', 'iphop_host_order', 'iphop_host_family',
'iphop_host_genus',
```

```
'PhaGCN_viral_phylum', 'PhaGCN_viral_class', 'PhaGCN_viral_order',
'PhaGCN_viral_family',
'taxmyphage_viral_Kingdom', 'taxmyphage_viral_Phylum', 'taxmyphage_viral_Class',
'taxmyphage_viral_Order', 'taxmyphage_viral_Family', 'taxmyphage_viral_Subfamily',
'taxmyphage_viral_Genus', 'taxmyphage_viral_Species'
```

Percentage of available viral/host taxonomy. Since there are two source of viral taxonomy (taxmyphage/PhaGCN), I use my curated metadata for model training and testing, which takes the value from taxmyphage when taxmyphage is available and takes the value from phaGCN when taxmyphage is not available.

```
Species is nan (taxmyphage): 0.9978568877247291 (18/8399 phages have this)
Genus is nan (taxmyphage): 0.9978568877247291
Sub-family is nan (taxmyphage): 0.9978568877247291
Family is nan (taxmyphage): 0.9978568877247291
Order is nan (taxmyphage): 0.9978568877247291
Class is nan (taxmyphage): 0.9978568877247291
Phylum is nan (taxmyphage): 0.9978568877247291
Kingdom is nan (taxmyphage): 0.9978568877247291

Family is nan (PhaGCN): 0.41338254554113585
Order is nan (PhaGCN): 1.0
Class is nan (PhaGCN): 0.41338254554113585
Phylum is nan (PhaGCN): 0.41338254554113585

Host genus is nan (iPHoP): 0.6085248243838552
Host family is nan (iPHoP): 0.6091201333492082
Host order is nan (iPHoP): 0.6085248243838552
Host class is nan (iPHoP): 0.6085248243838552
Host phylum is nan (iPHoP): 0.6085248243838552
```

STRAIN dataset

Sequences come from bacterial isolates; ground-truth hosts are known

Sequence lengths:

```
count      3239.000000
mean      31292.787589
std       18712.219444
min       10002.000000
25%       15445.000000
50%       27973.000000
75%       42848.000000
max      138160.000000
```

Metadata available:

```
 'contig_id', 'length', 'geNomad_viral_conservative', 'checkv_quality',
'iphop_host_confidence_score', 'iphop_host_phylum', 'iphop_host_class',
'iphop_host_order', 'iphop_host_family', 'iphop_host_genus',
'PhaGCN_viral_phylum', 'PhaGCN_viral_class', 'PhaGCN_viral_order',
'PhaGCN_viral_family', 'taxmyphage_viral_Kingdom',
'taxmyphage_viral_Phylum', 'taxmyphage_viral_Class',
'taxmyphage_viral_Order', 'taxmyphage_viral_Family',
'taxmyphage_viral_Subfamily', 'taxmyphage_viral_Genus',
'taxmyphage_viral_Species', 'ground_truth_host_phylum',
'ground_truth_host_class', 'ground_truth_host_order',
'ground_truth_host_family', 'ground_truth_host_genus',
'ground_truth_host_species'
```

Percentage of available viral/host taxonomy

```
iphop_host_genus is nan: 0.07286199444272924
iphop_host_family is nan: 0.07286199444272924
iphop_host_order is nan: 0.07286199444272924
iphop_host_class is nan: 0.07286199444272924
iphop_host_phylum is nan: 0.07286199444272924

ground_truth_host_species is nan: 0.0
ground_truth_host_genus is nan: 0.0
ground_truth_host_family is nan: 0.0
ground_truth_host_order is nan: 0.0
ground_truth_host_class is nan: 0.0
ground_truth_host_phylum is nan: 0.0

PhaGCN_viral_family is nan: 0.32201296696511267
PhaGCN_viral_order is nan: 1.0
PhaGCN_viral_class is nan: 0.32201296696511267
PhaGCN_viral_phylum is nan: 0.32201296696511267

taxmyphage_viral_Species is nan: 1.0
taxmyphage_viral_Genus is nan: 1.0
taxmyphage_viral_Subfamily is nan: 1.0
taxmyphage_viral_Family is nan: 1.0
taxmyphage_viral_Order is nan: 1.0
taxmyphage_viral_Class is nan: 1.0
taxmyphage_viral_Phylum is nan: 1.0
taxmyphage_viral_Kingdom is nan: 1.0
```

the matching between iPHoP host predictions and ground truth hosts (NaN dropped):

```
Phylum match: 99.97%
Class match: 99.97%
Order match: 99.53%
Family match: 76.96%
Genus match: 72.23%
```

MLP model

Implementation:

- Take the embeddings as input and predict viral/host taxonomy. Trained on Millardlab dataset (90% for training and 10% for validation)
- Experiment with two different training strategies: 1) predict for the whole test set (MATRIX or STRAIN); 2) split the test set in two halves and give one half to the training

Observation:

- Overfitting is generally observed. While the training and validation accuracies are high, the testing accuracies are not as good as that.
- The viral taxonomy can be tricky. In the test sets (MATRIX & STRAIN), We lack Genus/Subfamily/Order information for most genomes, and we only have one Class/Phylum, which means that only the results on Family prediction are more informative. How well the model learns about each family can be divergent.
- The splitting strategy significantly made the results better.
- Host taxonomy prediction works significantly better with the STRAIN dataset where ground truth hosts are known.

Results of training and testing

The detailed report can be found in 'documents/model_comparison.md'. More plots (i.e. plots for training) can be found in the folder 'documents/img'

Trained on Millard only; Tested on MATRIX

```
Train Accuracies: {'Genus': '0.7801', 'Sub-family': '0.9440', 'Family': '0.9556',
'Order': '0.9969', 'Class': '0.9622', 'Phylum': '0.9620', 'Kingdom': '0.9621',
'Realm': '0.9621', 'Host_phylum': '0.9879', 'Host_class': '0.9825', 'Host_order':
'0.9545', 'Host_family': '0.9263', 'Host_genus': '0.8698'}
Val Accuracies: {'Genus': '0.8211', 'Sub-family': '0.9599', 'Family': '0.9692',
'Order': '0.9990', 'Class': '0.9617', 'Phylum': '0.9620', 'Kingdom': '0.9617',
'Realm': '0.9617', 'Host_phylum': '0.9892', 'Host_class': '0.9829', 'Host_order':
'0.9475', 'Host_family': '0.9260', 'Host_genus': '0.8750'}
```

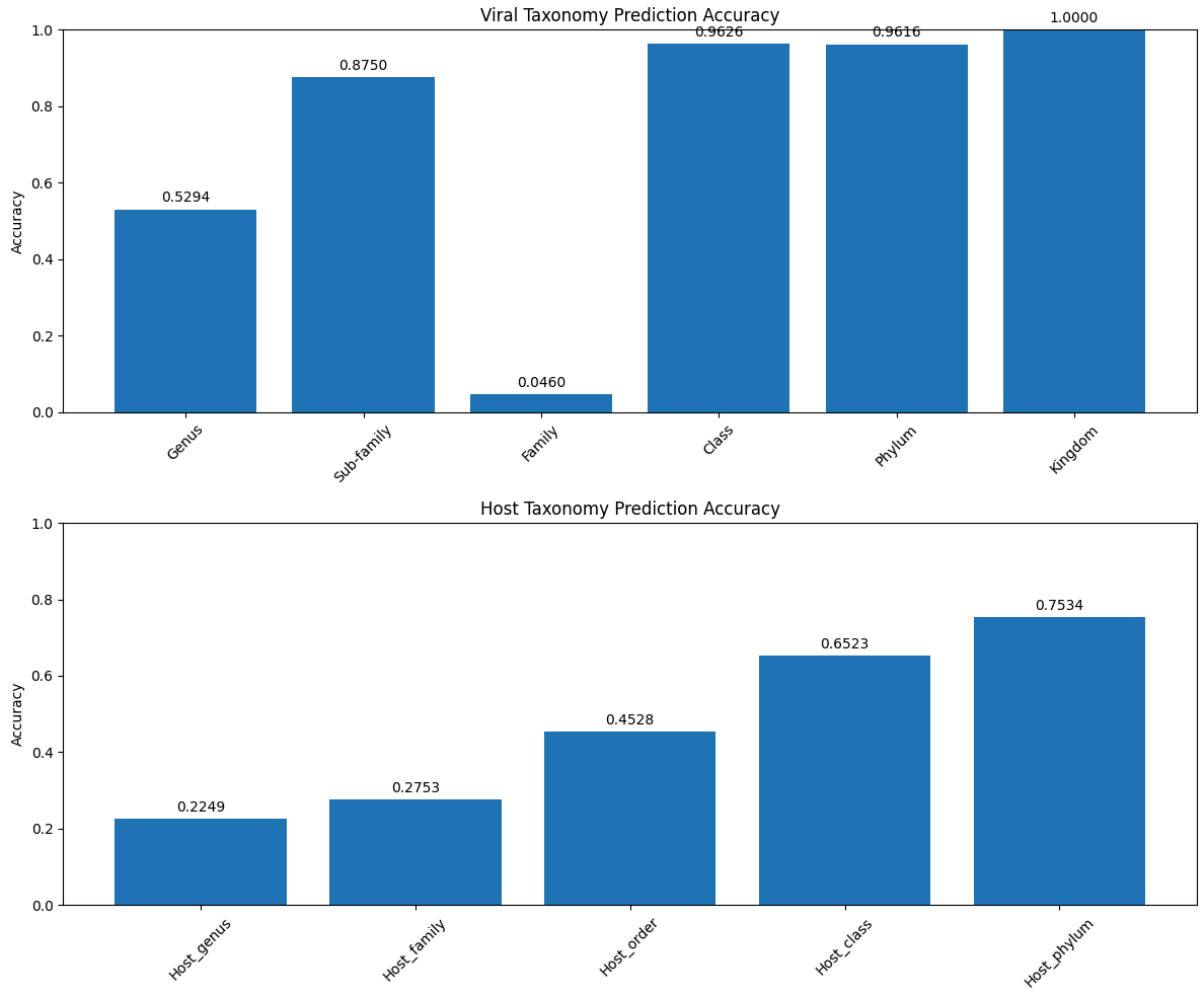

Figure: train_millard-test_matrix

Trained on Millard & MATRIX (half); Tested on MATRIX (another half)

```
Train Accuracies: {'Genus': '0.8814', 'Sub-family': '0.9855', 'Family': '0.9065',
'Order': '0.9950', 'Class': '0.9981', 'Phylum': '0.9981', 'Kingdom': '0.9984',
'Host_genus': '0.8463', 'Host_family': '0.9043', 'Host_order': '0.9445',
'Host_class': '0.9771', 'Host_phylum': '0.9857'}
Val Accuracies: {'Genus': '0.9044', 'Sub-family': '0.9854', 'Family': '0.9003',
'Order': '0.9942', 'Class': '0.9975', 'Phylum': '0.9975', 'Kingdom': '0.9970',
'Host_genus': '0.8509', 'Host_family': '0.9072', 'Host_order': '0.9398',
'Host_class': '0.9734', 'Host_phylum': '0.9856'}
```

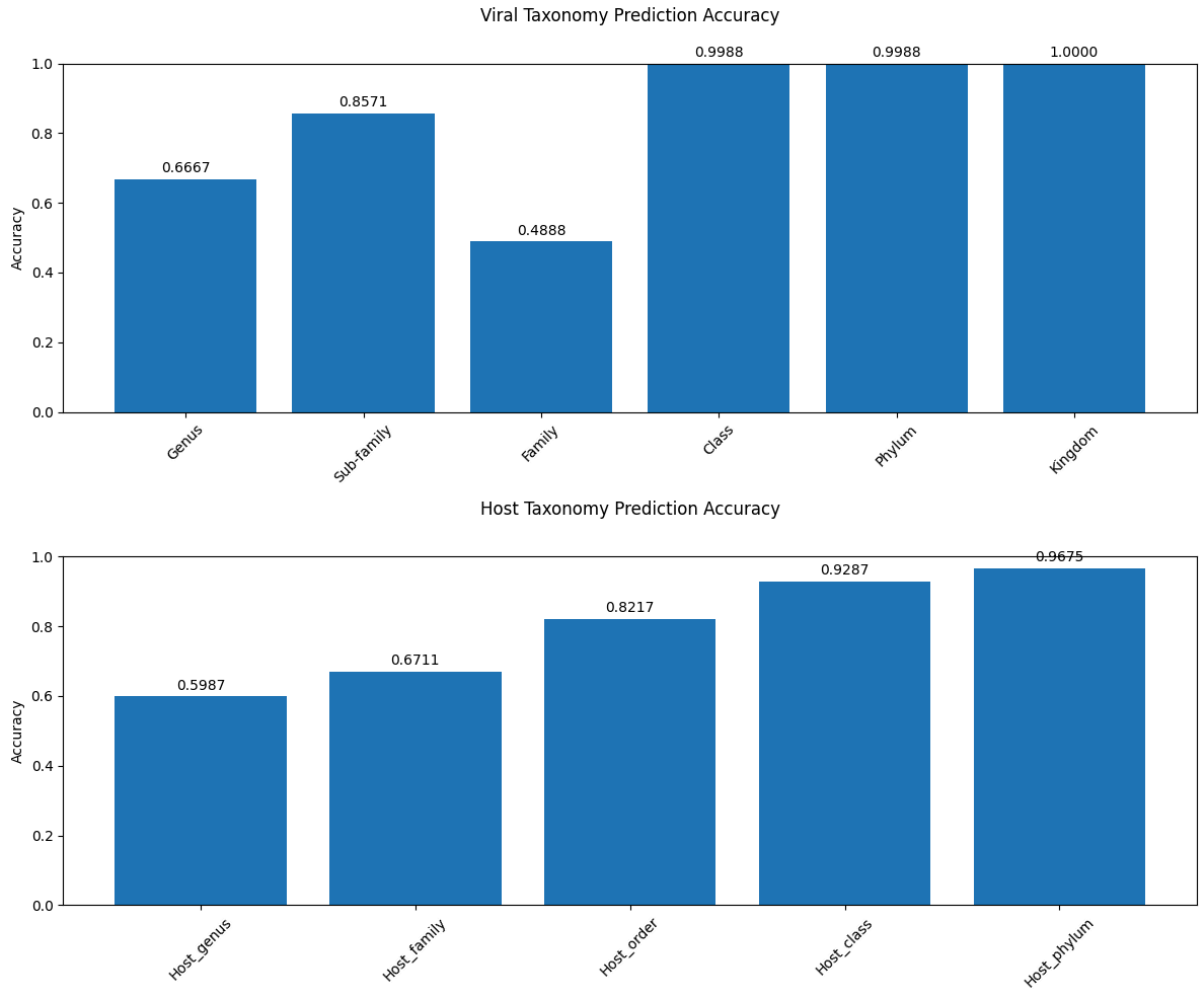

Figure: train_millard&matrix-test_matrix

Trained on Millard only; Tested on STRAIN

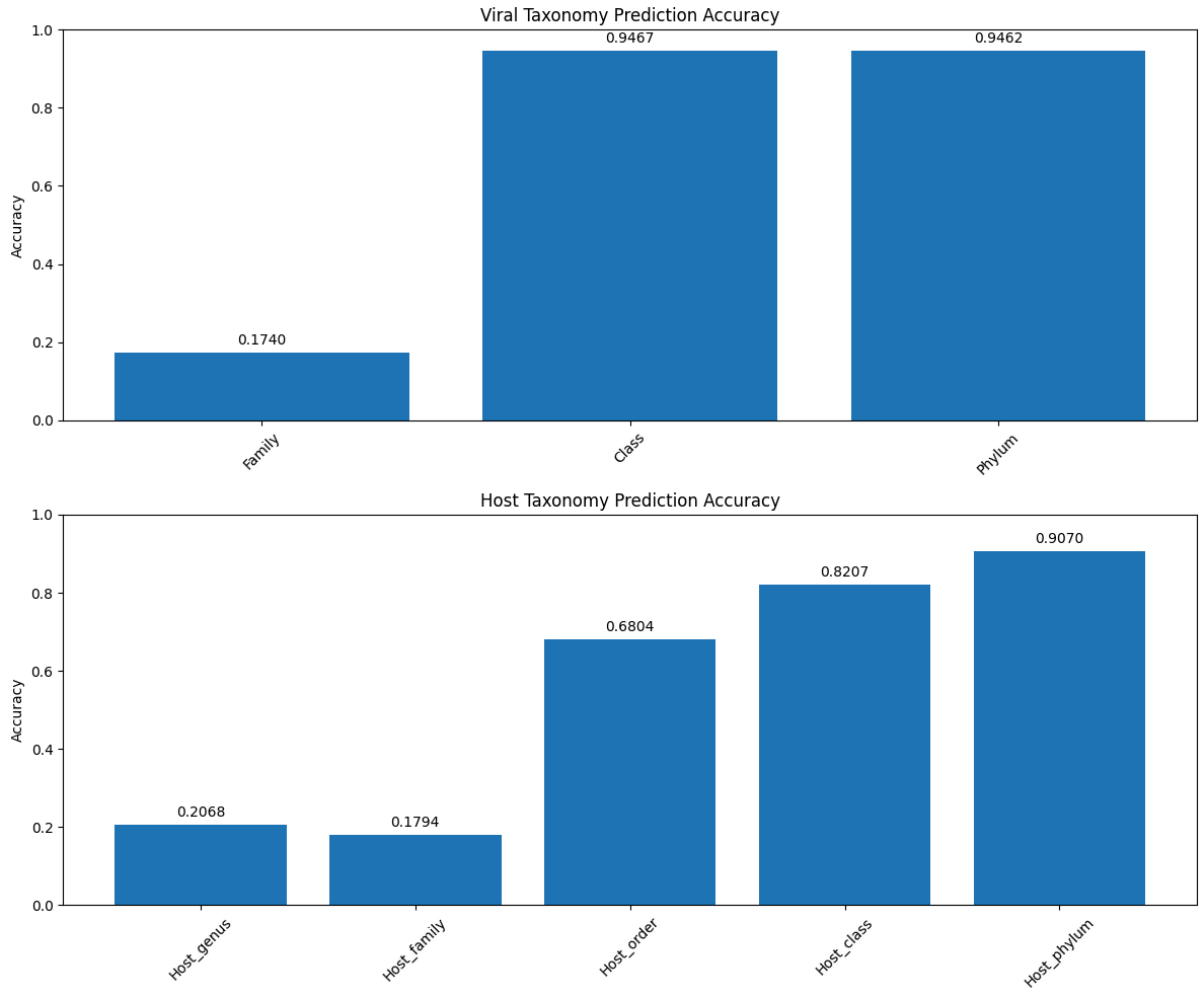

Figure: train_millard-test_strain

Trained on Millard & STRAIN (half); Tested on STRAIN (another half)

```
Train Accuracies: {'Family': '0.9851', 'Order': '0.9987', 'Class': '0.9991',
'Phylum': '0.9991', 'Host_genus': '0.8733', 'Host_family': '0.9277', 'Host_order':
'0.9579', 'Host_class': '0.9826', 'Host_phylum': '0.9889'}
Val Accuracies: {'Family': '0.9840', 'Order': '0.9953', 'Class': '0.9979',
'Phylum': '0.9975', 'Host_genus': '0.8706', 'Host_family': '0.9253', 'Host_order':
'0.9567', 'Host_class': '0.9793', 'Host_phylum': '0.9885'}
```

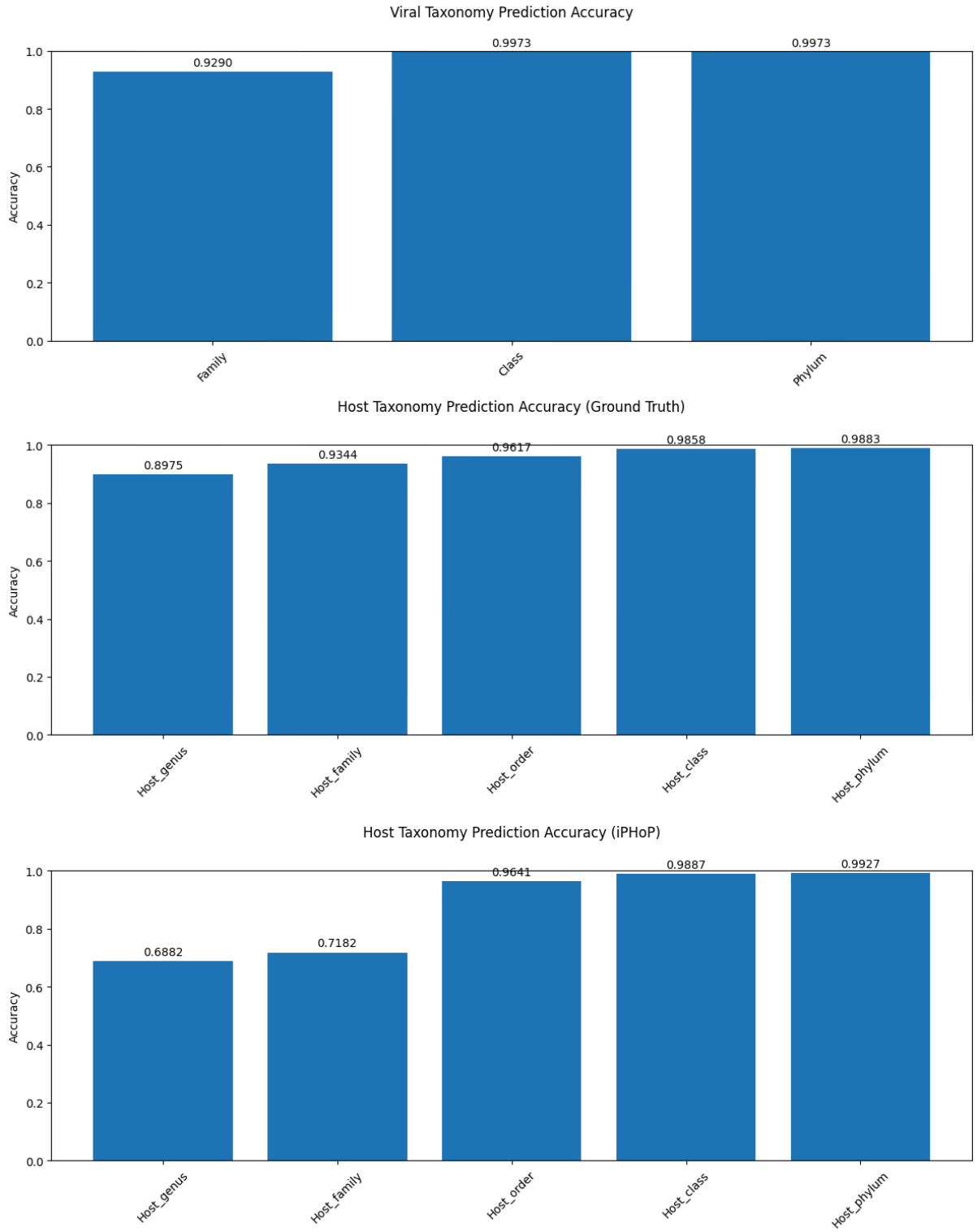

Figure: train_millard&strain-test_strain

We can note that the results for iPHoP are not as good as Ground-Truth when host genera or host families are predicted. This is likely due to the mismatch of iPHoP and ground-truth host information (which can be checked in the ‘dataset’ section in this report). When order/class/phylum are predicted, the result for iPHoP excels because iPHoP have incomplete host taxonomies and some samples that lower the accuracies for ground-truth are omitted when we test for iPHoP (which can be checked in the full report).