Novelty detection using NBC

Overview

- Used 5 different k-mer lengths (3, 6, 9, 12 and 15)
- Taxonomic levels: phylum, class, order and family
- 5 trials for each k-mer length at each level, 100 trials in total

Creating the training data for each trial

- Database consisted of 4634 unique species, with 1 genome per species
- Different taxa had varying number of representation in the database
- Eg. Bacillota had 325 representatives while some of them just have 1
- Excluded all taxa with less than 30 unique representatives
- Remaining taxa was then randomly sampled to create training sets

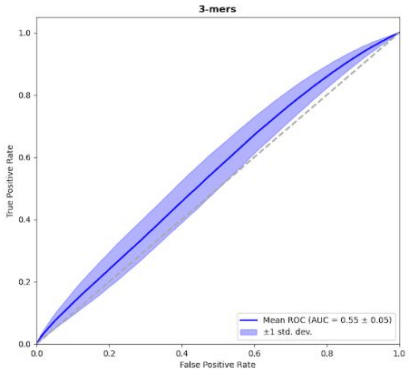
	Phylum
Uroviricota	1139
Pseudomonadota	762
Lenarviricota	443
Bacillota	325
Actinomycetota	246
...	...
Calditrichota	1
Lentisphaerota	1
Balneolota	1
Nitrospinota	1
Cossaviricota	1

Results

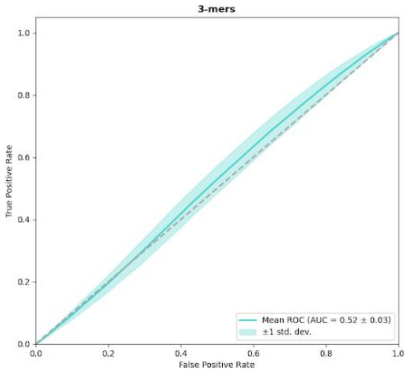
- All genomes in the training data were labeled as ‘known’ and the rest outside of the training were labeled ‘unknown’
- Plotted ROCs for each trial

3-mers

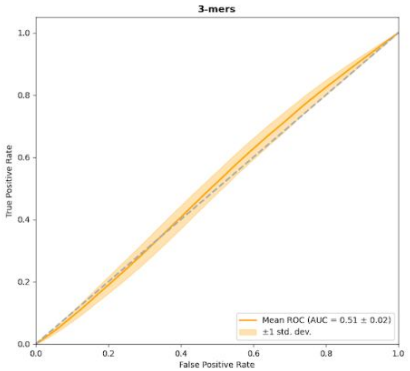
Phylum



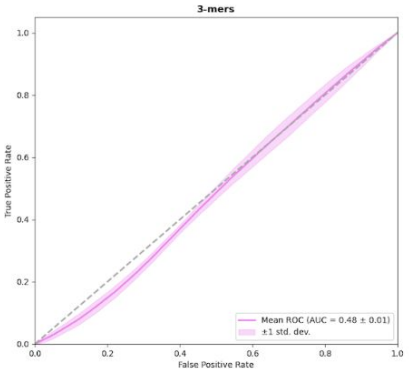
Class



Order

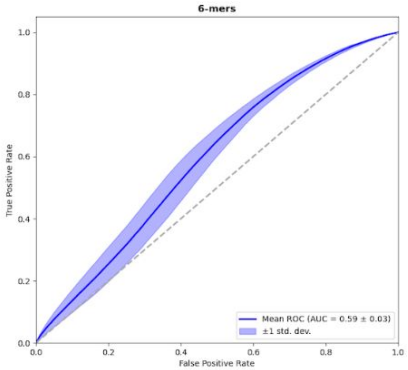


Family

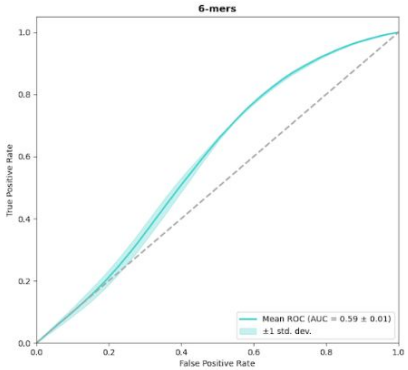


6-mers

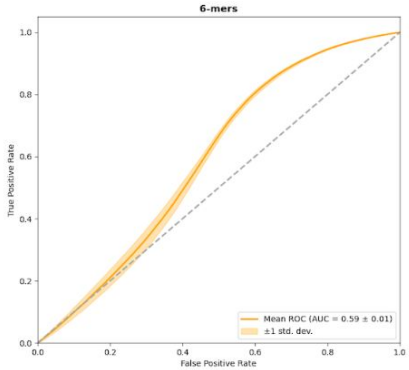
Phylum



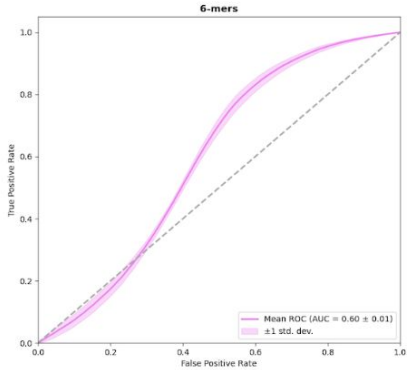
Class



Order

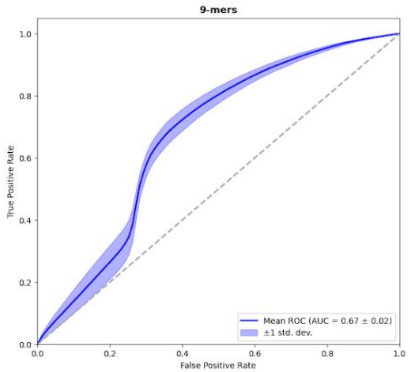


Family

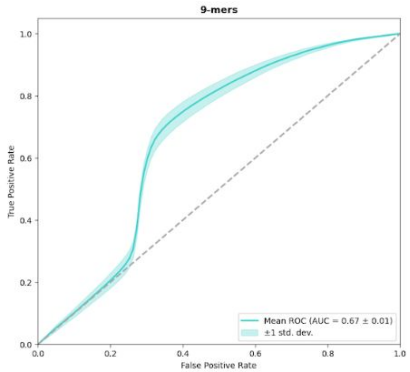


9-mers

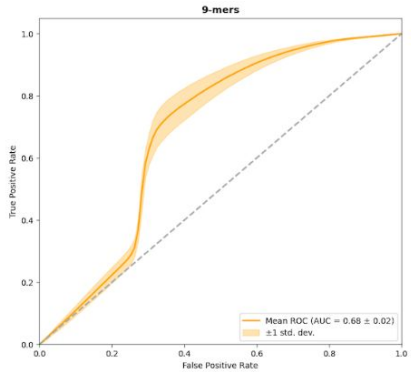
Phylum



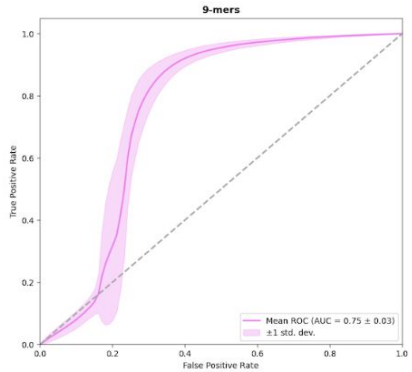
Class



Order

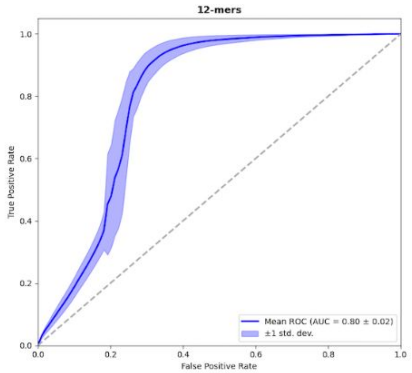


Family

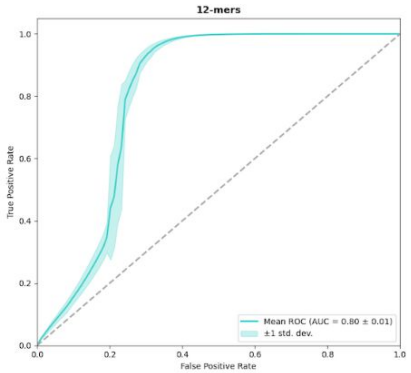


12-mers

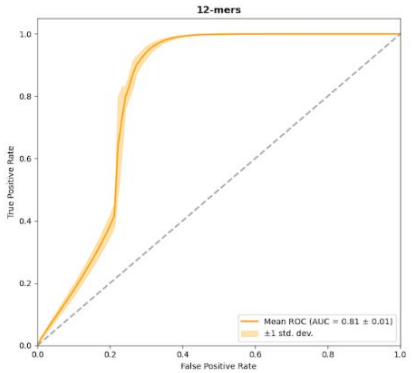
Phylum



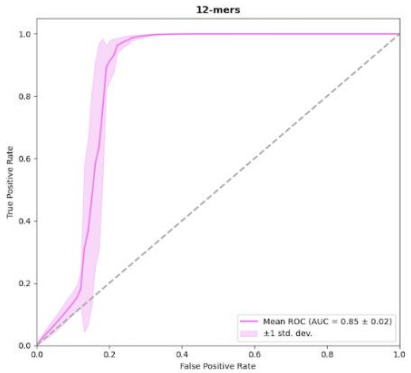
Class



Order

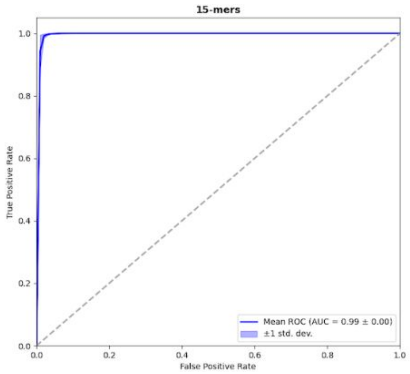


Family

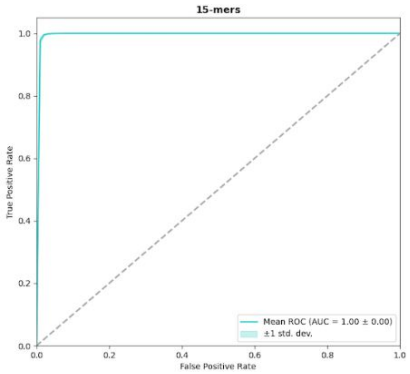


15-mers

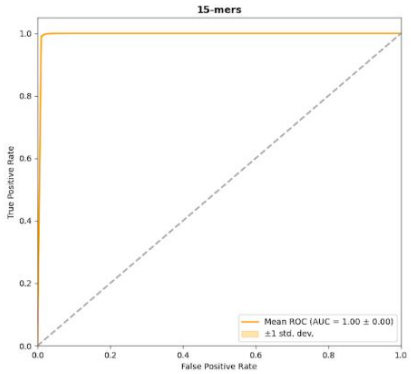
Phylum



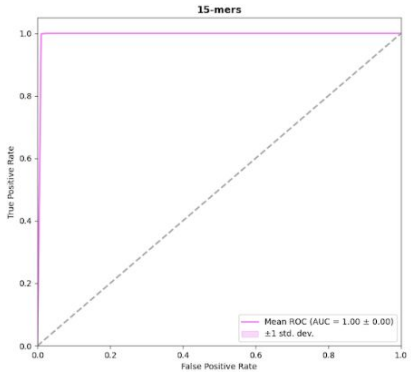
Class



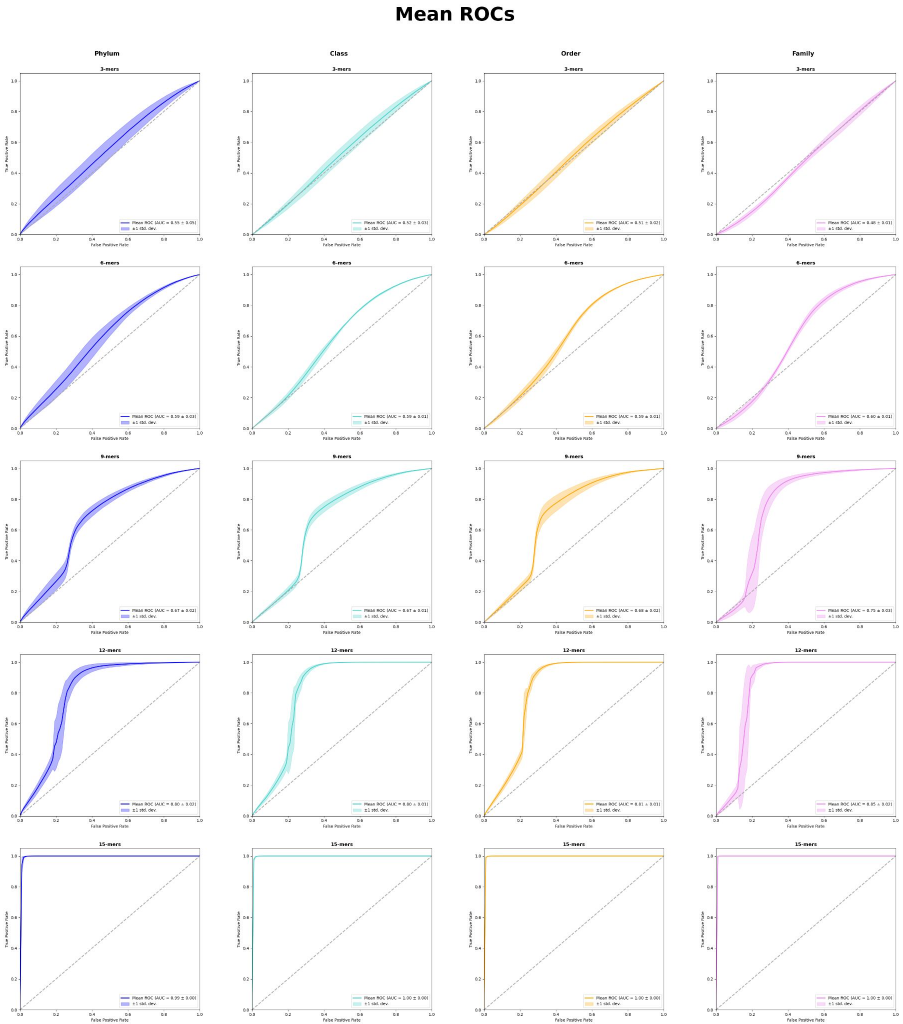
Order



Family



Mean ROC/AUC


Mean ROCs
Phylum
Class
Order
Family
3-mers
6-mers
9-mers
12-mers
15-mers