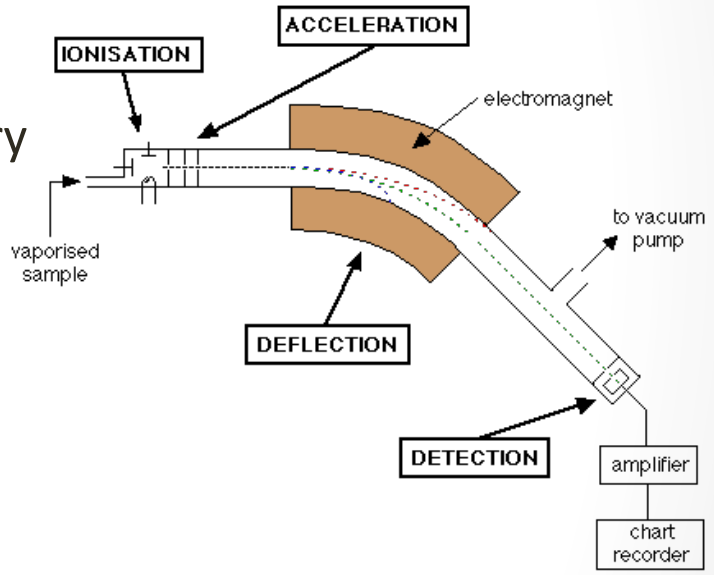
Random Forest Classification for Identifying Bacteria with Mass Spectrometry in Mixed Samples

David Gray

July 1, 2019

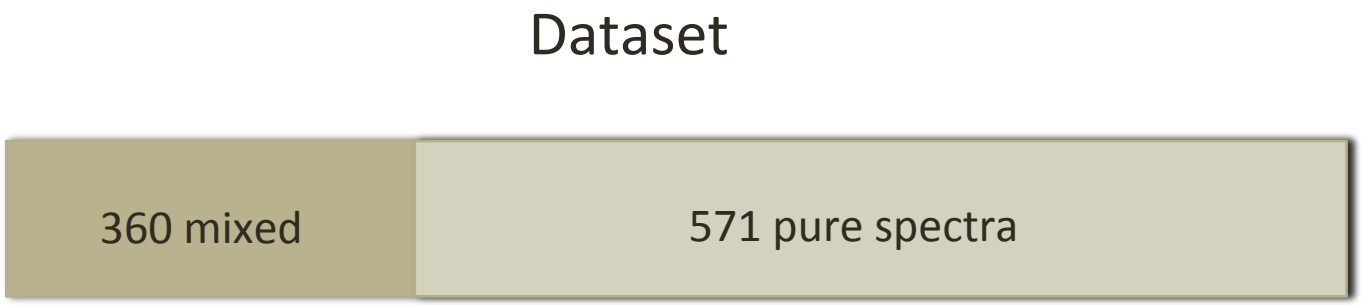
Machine Learning for Microorganism Identification

- The dataset used in this study is described at the UCI machine learning repository as “a dataset to explore machine learning approaches for the identification of microorganisms from mass-spectrometry data.”
- MS-based microbial identification is not feasible routinely
- One of the issues is that polymicrobial samples can will yield a ‘mixed’ MS fingerprint.
- The purpose of this work is to use machine learning to try to decipher identity from samples with two species of bacteria.

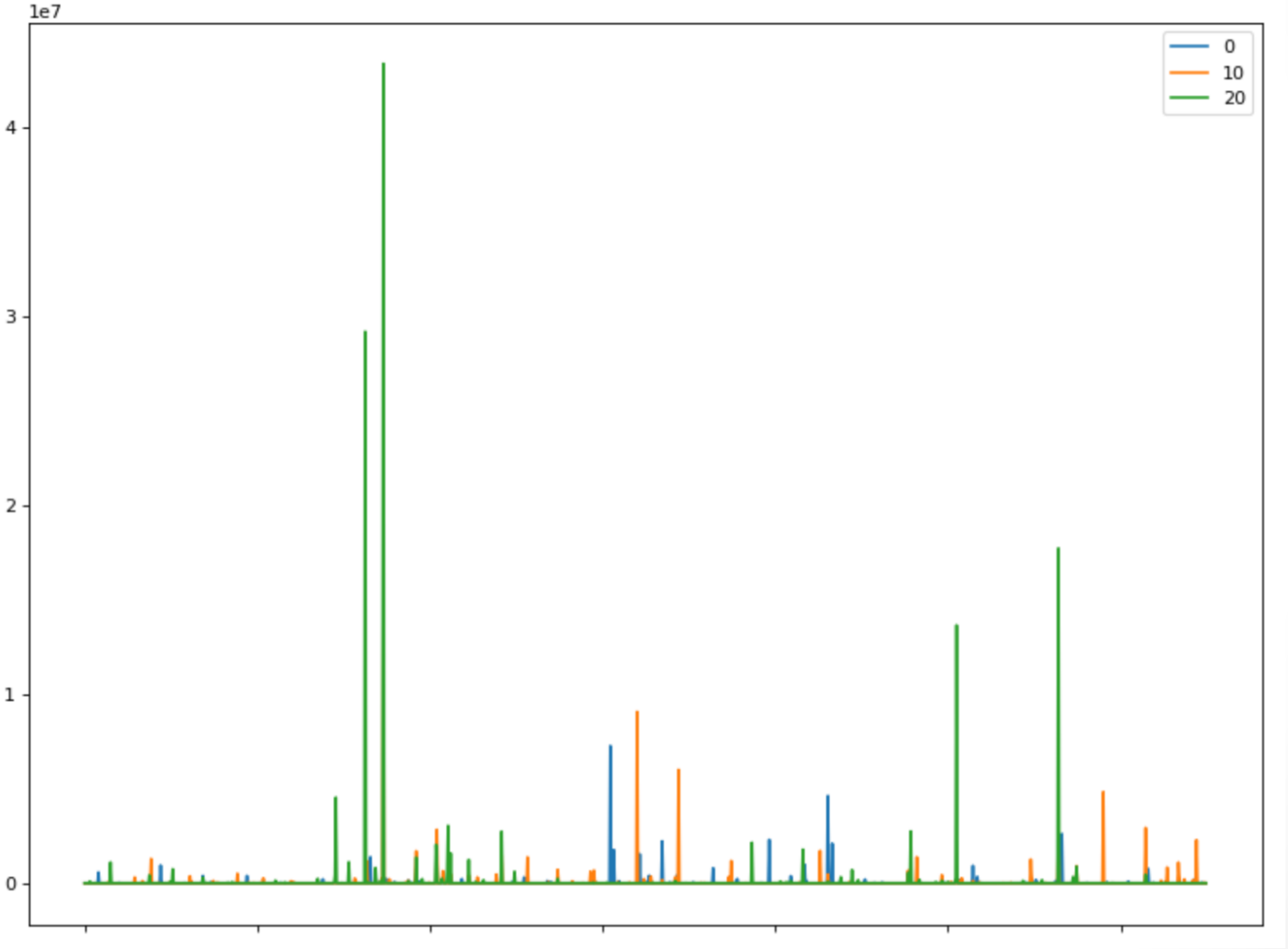


Data Set Constituents

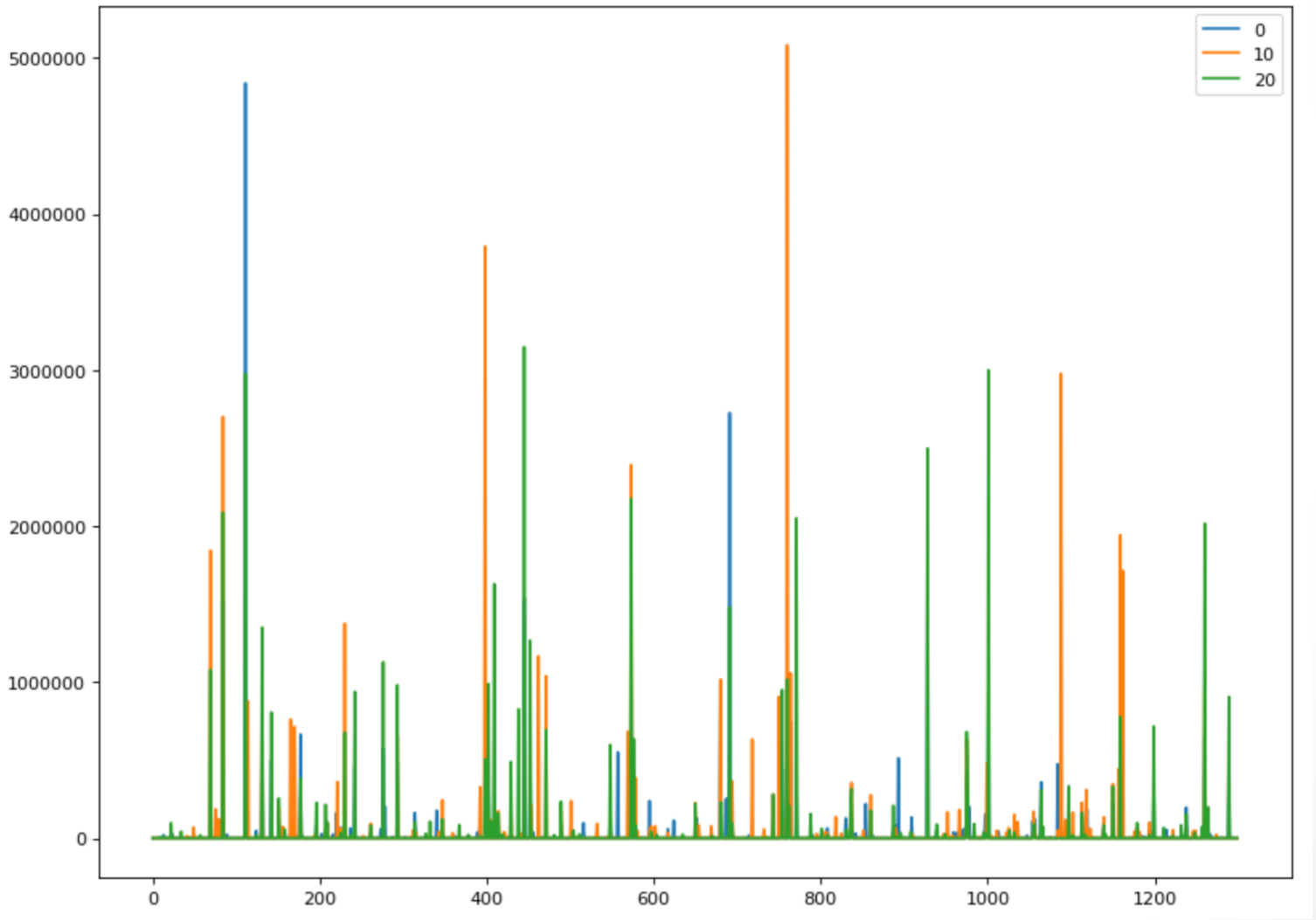
- The dataset includes 931 MALDI-TOF mass spectra
- 571 spectra of pure, individual samples
- The “mixed dataset” consists of 360 spectra



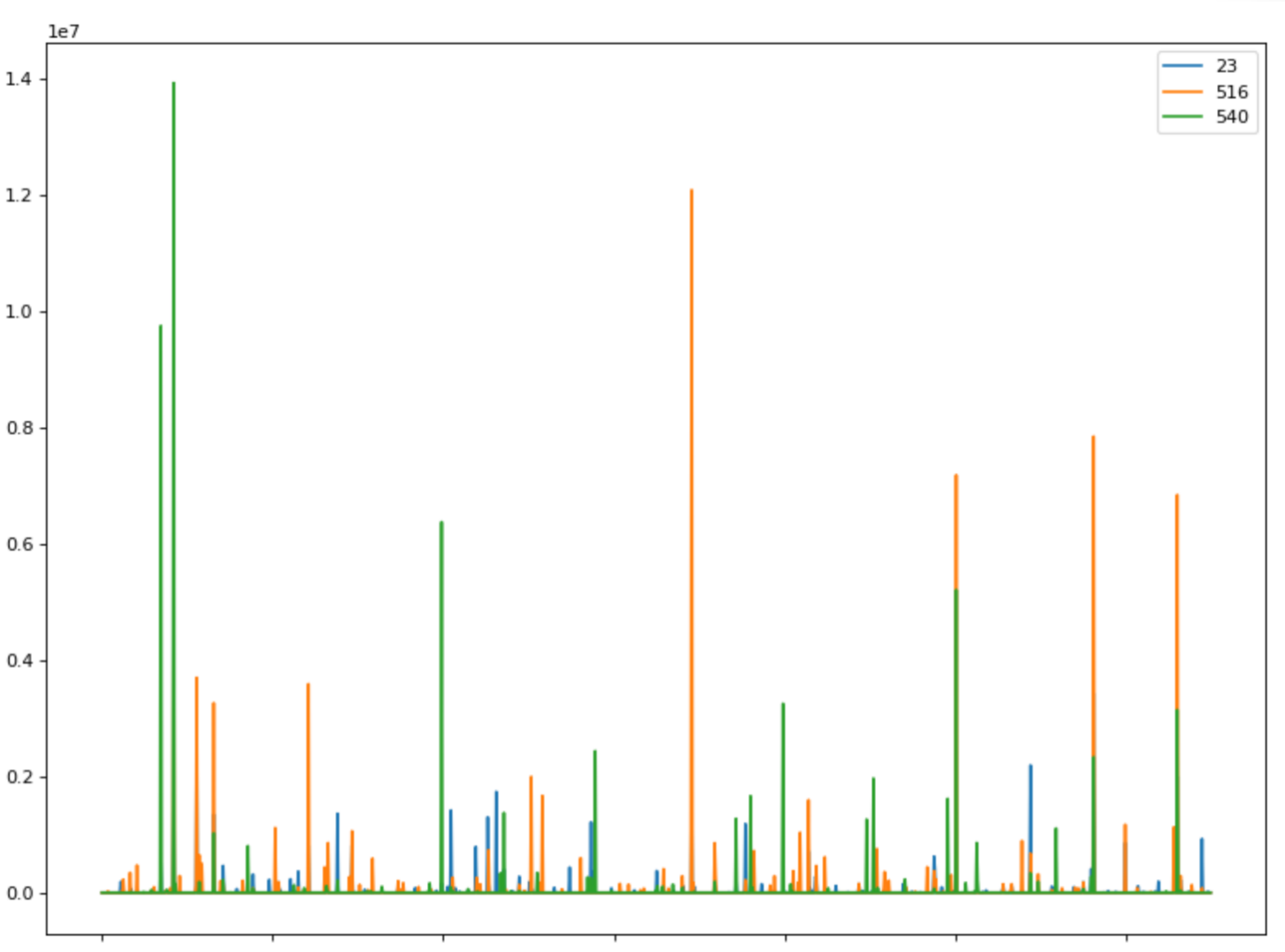
Pure Spectra of Samples 0, 10, and 20



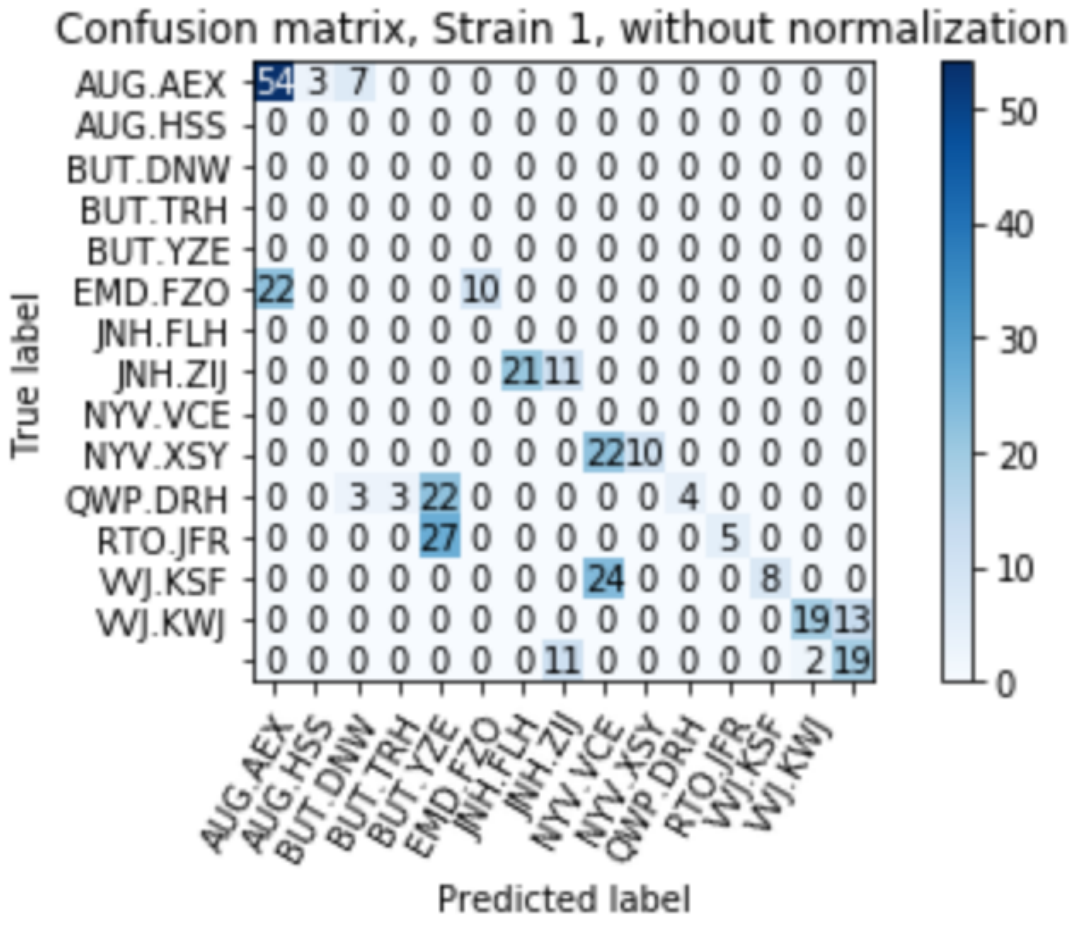
Mixed Spectra of Samples 0, 10, 20



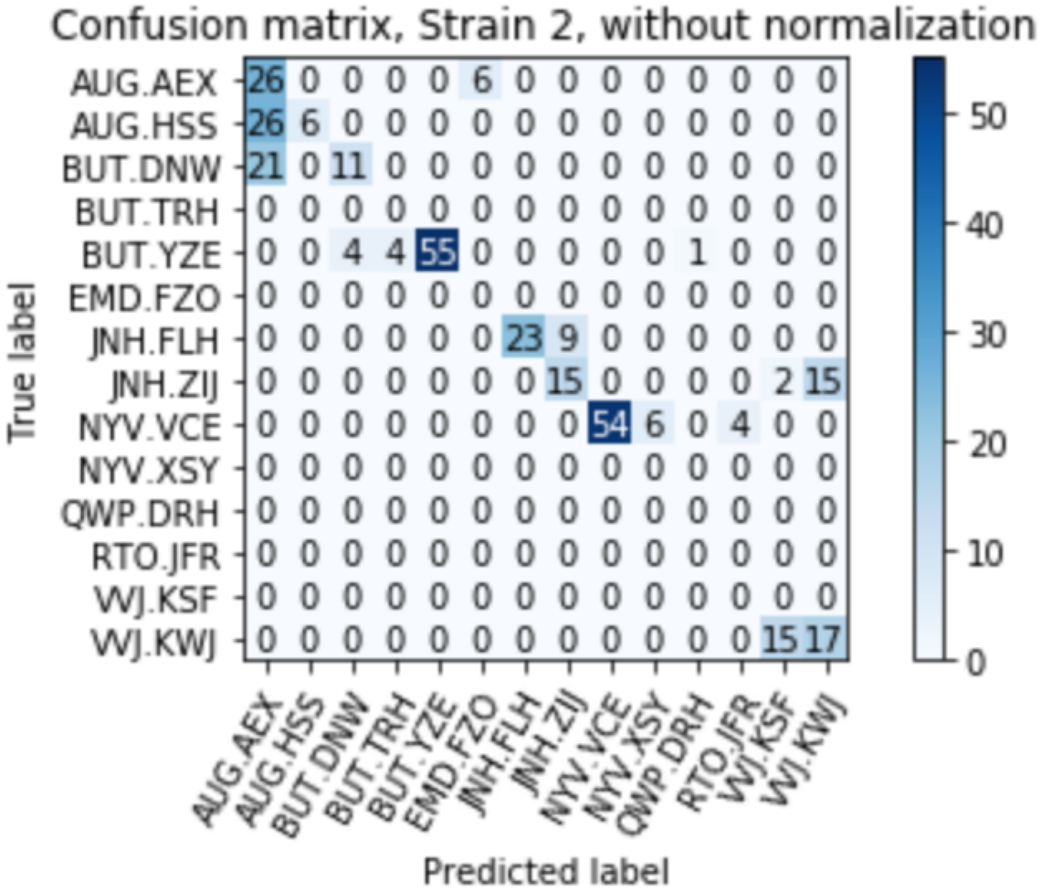
Mixed and Pure Spectra



Confusion Matrix Predicting Strain 1



Confusion Matrix Predicting Strain 2



Accuracy of Prediction and Summary

Accuracy of Predicting Strain 1	**43.8%**
Accuracy of Predicting Strain 2	**64.7%**

- A few strains were particularly well predicted, especially from Strain 2 samples.
- The authors of the paper pioneering analysis on this data and simulated data reported: "Few spectra were misidentified and mixtures were always at least partially identified. More than 60% of the mixtures were detected and correctly identified."
- Our model would need to be modified to try to predict both bacteria types in the mixed samples as there is only one species in this work predicted per mixed sample.
- It will be interesting to compare the samples correctly predicted that are in a higher or lower dilution. Comparing the average correctly predicted with different techniques or parameters will help refine the method. It would be interesting to look at other statistics to evaluate the model beyond accuracy.