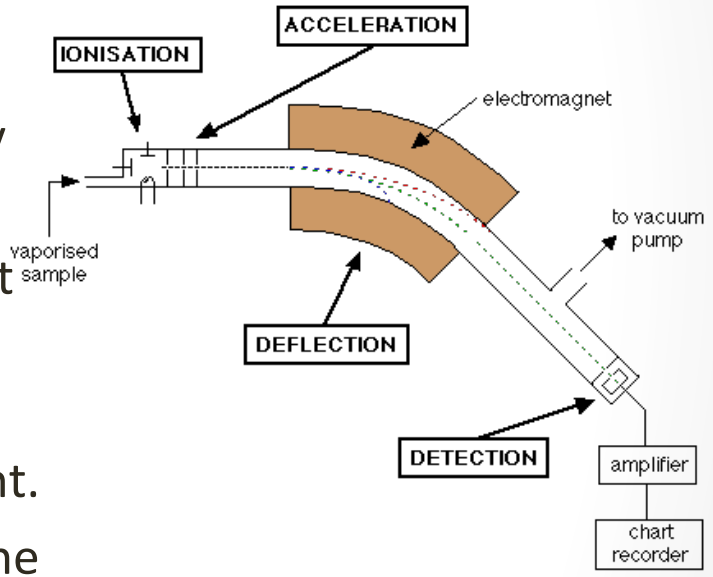
Random Forest Classification for Identifying Bacteria with Mass Spectrometry in Mixed Samples

David Gray

July 18, 2019

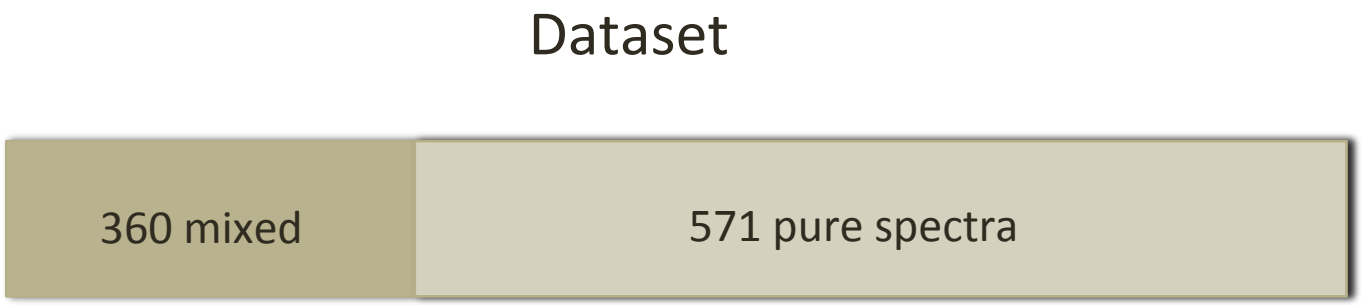
Machine Learning for Microorganism Identification

- 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 has not been feasible routinely
- One of the issues is that polymicrobial samples can 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.

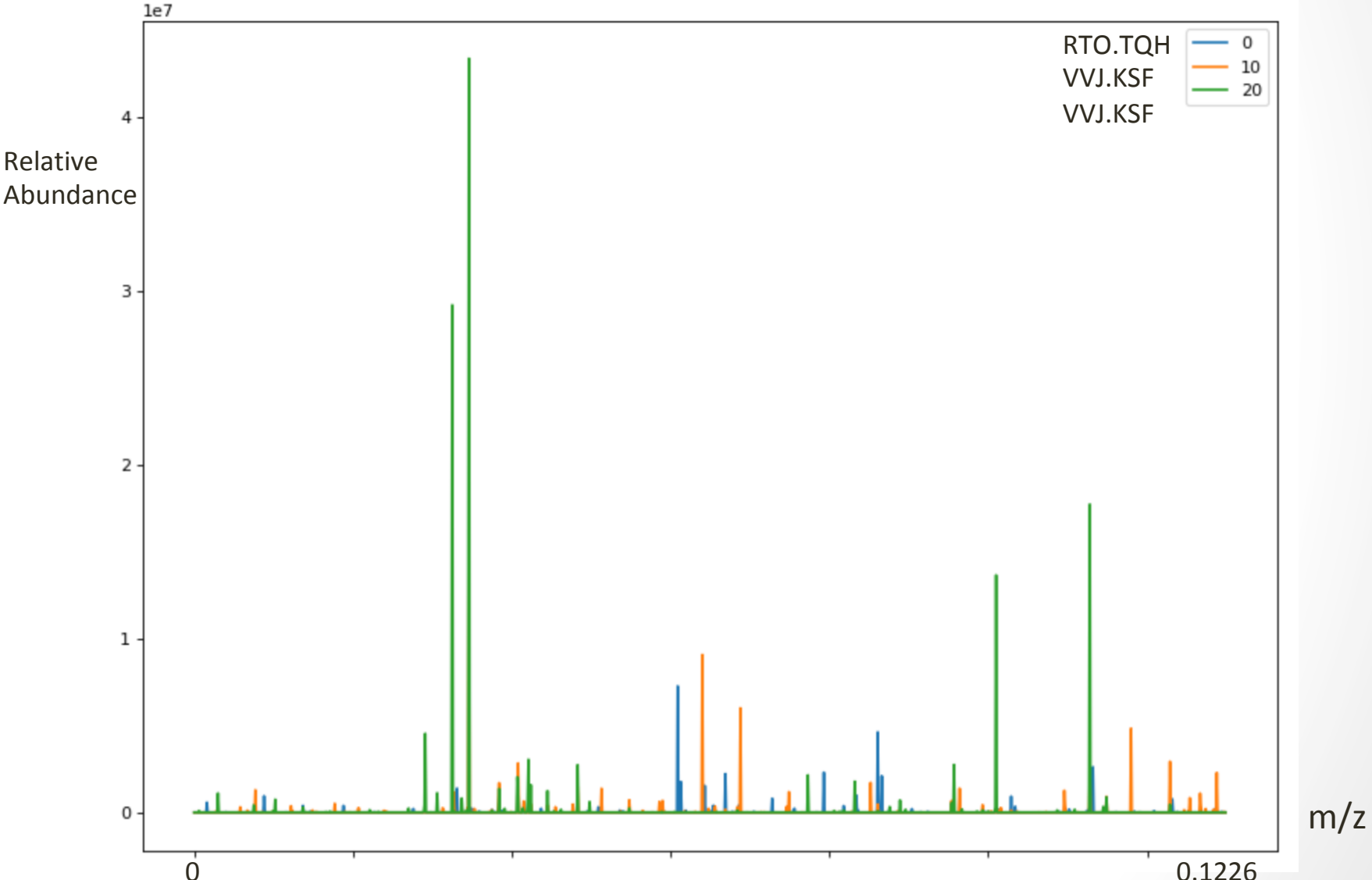


Data Set Constituents

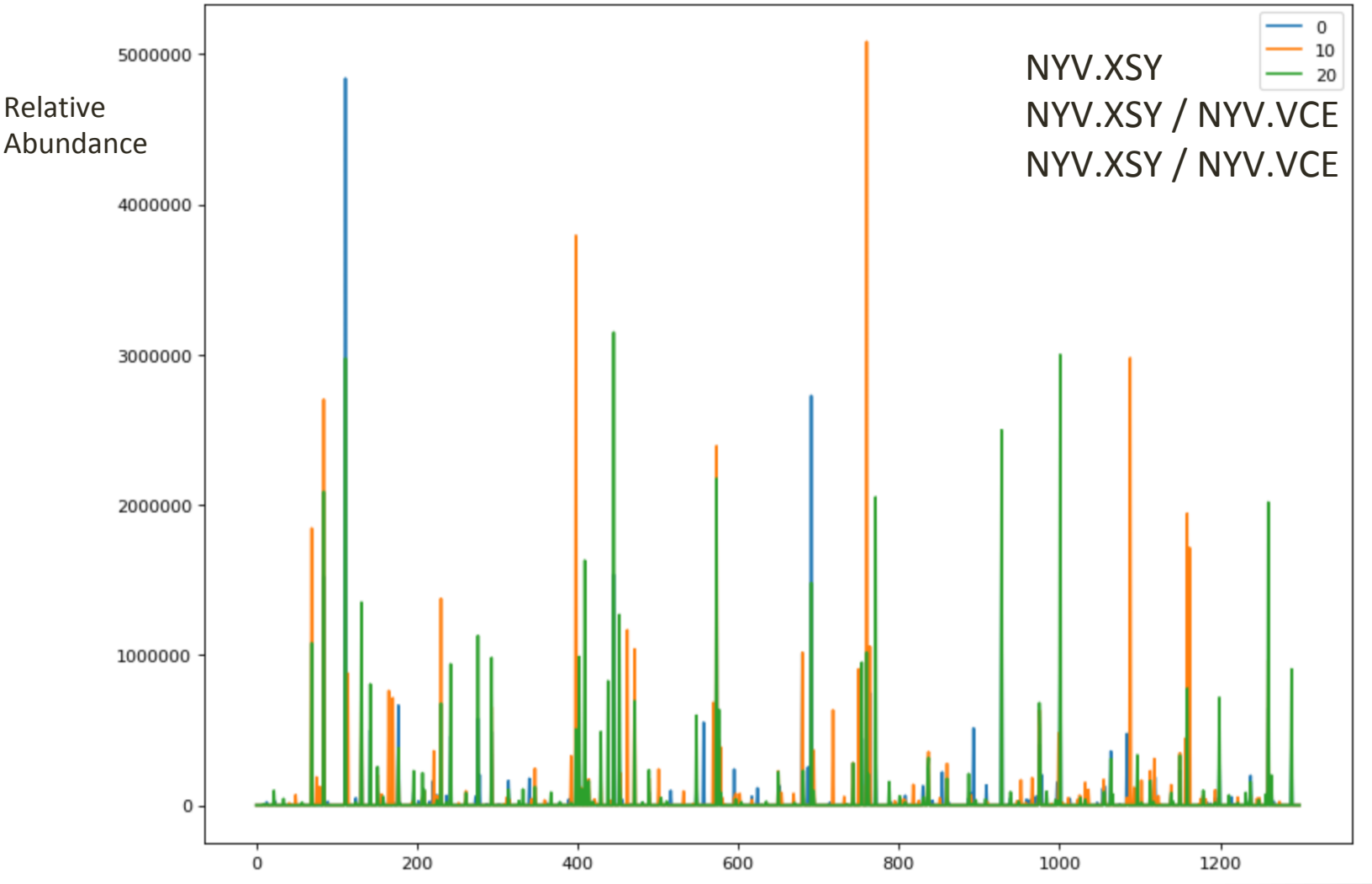
- The dataset includes 931 MALDI-TOF mass spectra
  - 571 spectra of pure, individual samples
  - The “mixed dataset” consists of 360 spectra



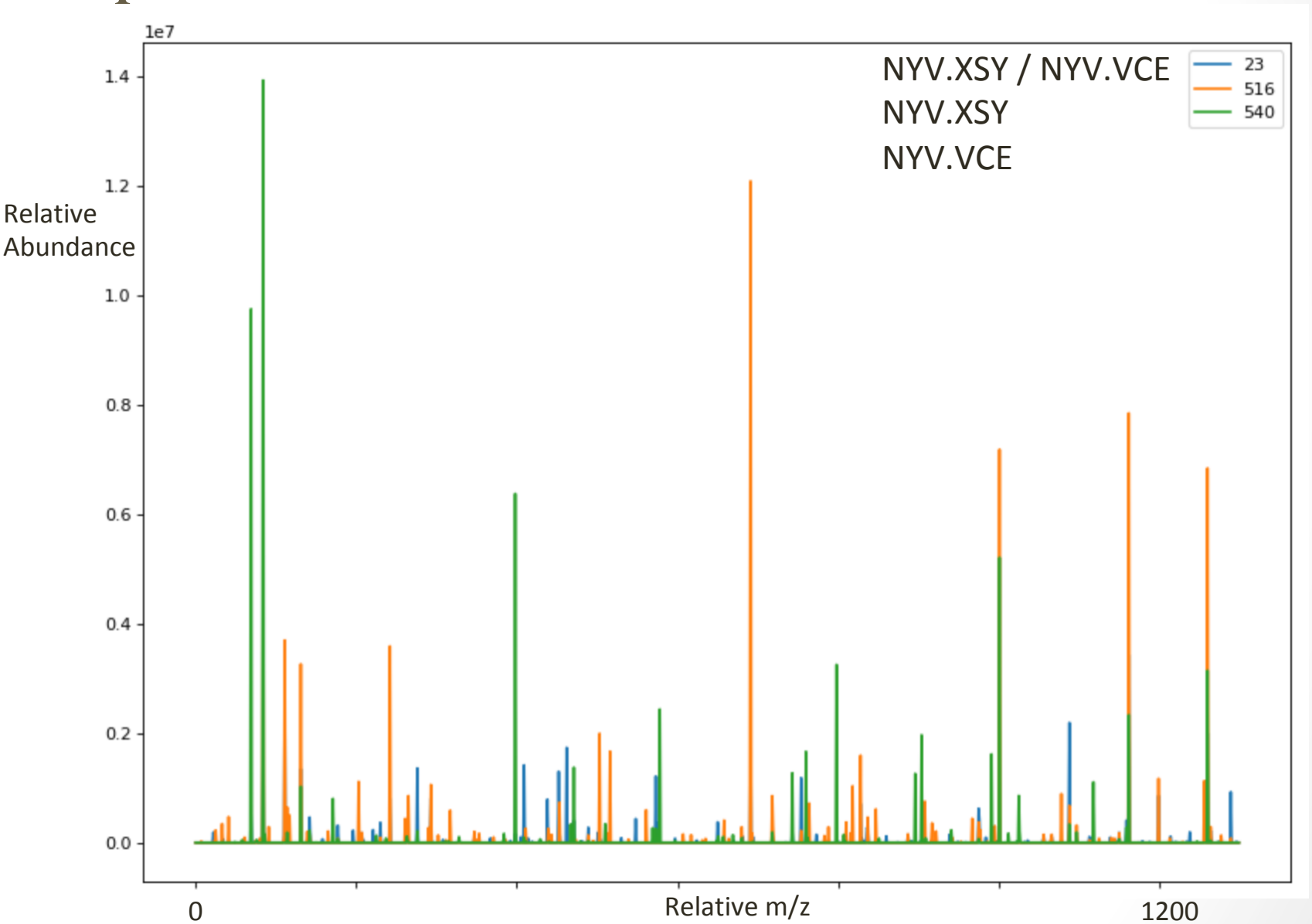
Pure Spectra of Samples 0, 10, and 20



Mixed spectra of samples (0 (one strain, same) and 10 and 20 (one different strains))



Mixed (23) and Pure (516 and 540) Spectra

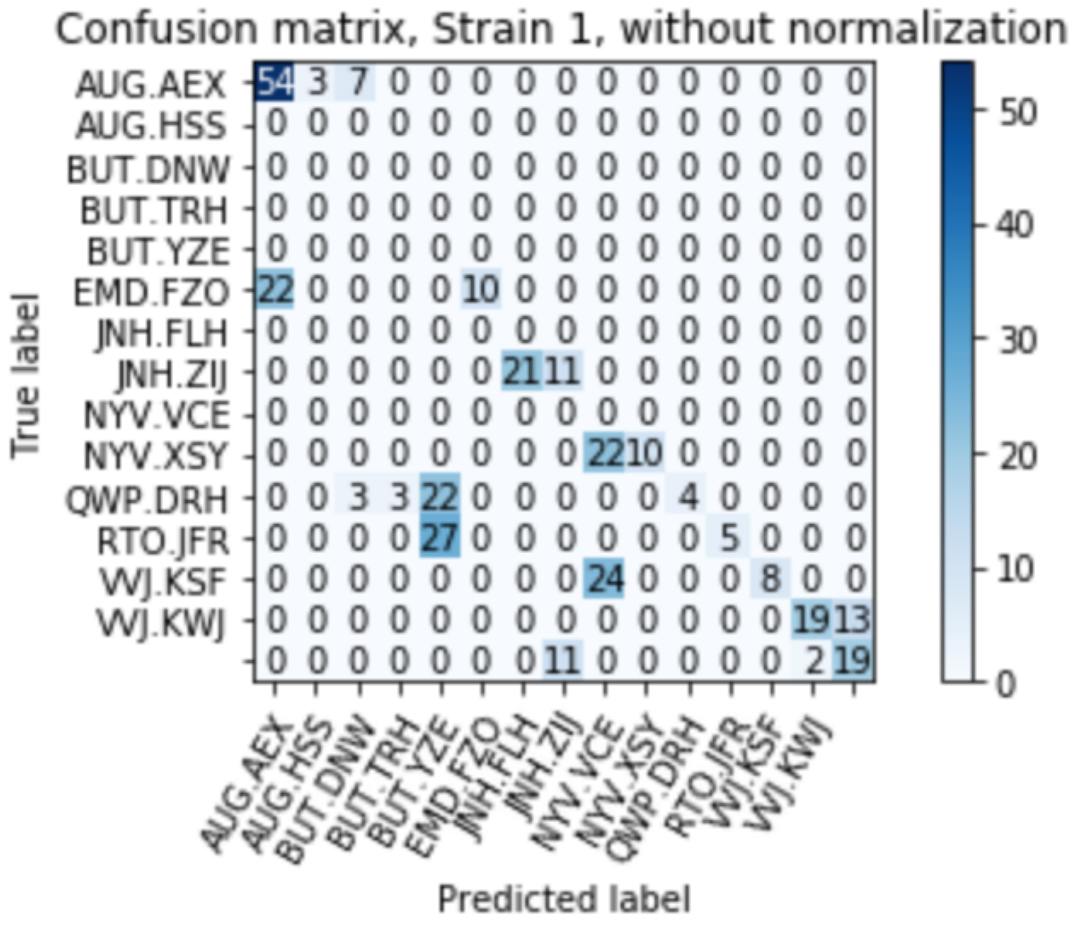


A Random Forest Classifier was Used to Detect Bacteria

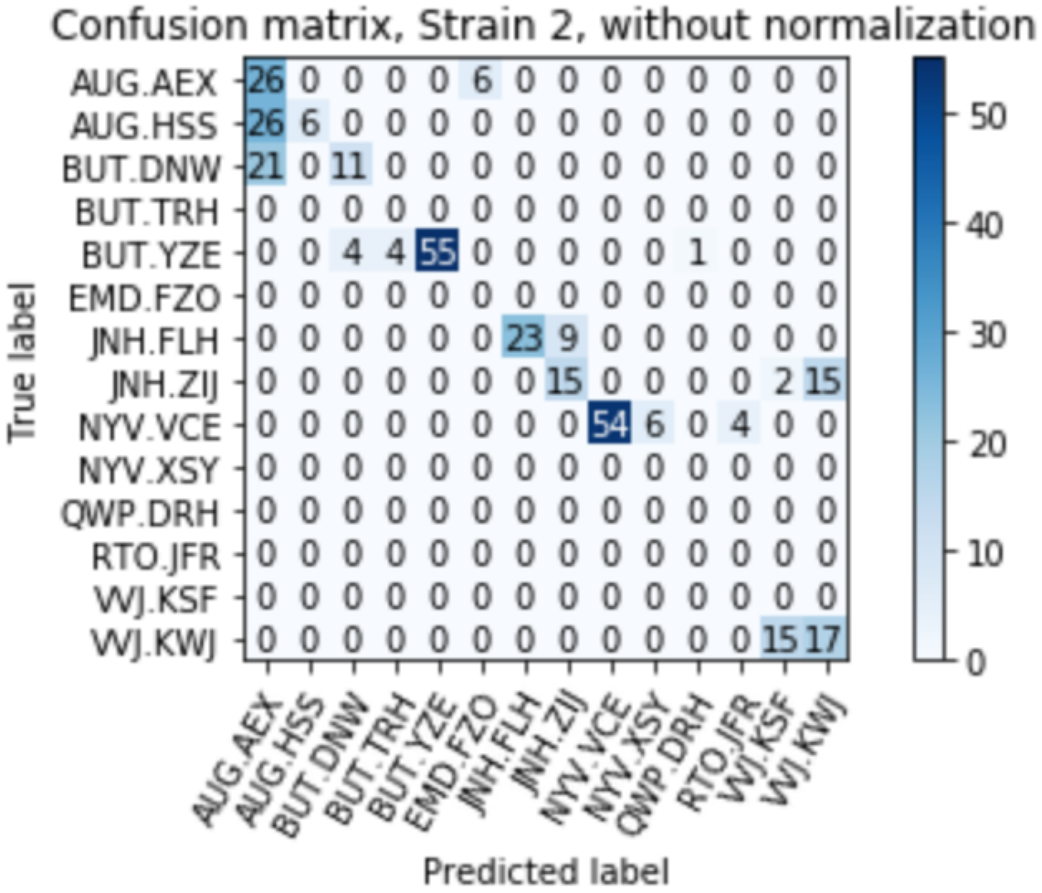
- This is a simple model although does not usually match the best supervised learning approaches in accuracy of prediction.
  - We used this model as a first step for prediction for learning
  - The work could be refined to employ a different classifier
- The classifier was trained on the pure spectra, and then it evaluated the mixed spectra to make one prediction

Parameter	Value
n_estimators	100
bootstrap	“True”
max_depth	None
min_samples_split	5
min_impurity_decrease	1e-5
max_features	“auto”

Confusion Matrix Predicting Strain 1: 43.8% Accuracy



Confusion Matrix Predicting Strain 2: 64.7% Accuracy



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 approach would need to be modified to try to predict both bacteria types in the mixed samples as there is only one species predicted in this work per mixed sample at a time
- It will be interesting to compare the samples correctly predicted that are in a higher or lower dilution. It would be worthwhile to count the number of pure samples predicted accurately
- 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.