

5-2022

Using a Machine Learning Model to Predict Plant Inflorescences based upon its Soil Microbiome

Luke Denoncourt

Follow this and additional works at: <https://scholarworks.wm.edu/honorstheses>



Part of the [Data Science Commons](#), [Environmental Microbiology and Microbial Ecology Commons](#), and the [Plant Sciences Commons](#)

Recommended Citation

Denoncourt, Luke, "Using a Machine Learning Model to Predict Plant Inflorescences based upon its Soil Microbiome" (2022). *Undergraduate Honors Theses*. William & Mary. Paper 1896.
<https://scholarworks.wm.edu/honorstheses/1896>

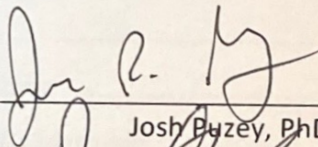
This Honors Thesis -- Open Access is brought to you for free and open access by the Theses, Dissertations, & Master Projects at W&M ScholarWorks. It has been accepted for inclusion in Undergraduate Honors Theses by an authorized administrator of W&M ScholarWorks. For more information, please contact scholarworks@wm.edu.

Using a Machine Learning Model to Predict Plant Inflorescences based upon its Soil
Microbiome

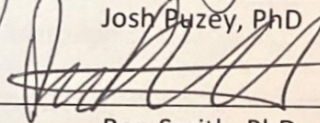
A thesis submitted in partial fulfillment of the requirement
for the degree of Bachelor of Arts / Science in Department from
William & Mary

by

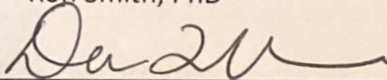
Luke Denoncourt

A handwritten signature in dark ink, appearing to read "Josh Puzey", written over a horizontal line.

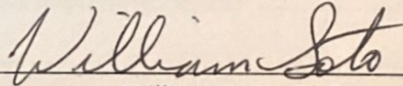
Josh Puzey, PhD

A handwritten signature in dark ink, appearing to read "Ron Smith", written over a horizontal line.

Ron Smith, PhD

A handwritten signature in dark ink, appearing to read "Dana Willner", written over a horizontal line.

Dana Willner, PhD

A handwritten signature in dark ink, appearing to read "William Soto", written over a horizontal line.

William Soto, PhD

Williamsburg, VA

May 5, 2022

Abstract

The UN estimates that the global population could reach 9.7 billion by 2050 (United Nations). As a result, the amount of food required to feed humanity is thought to double by 2050 (Ray et al., 2012). Humanity must find a way to increase crop production without increasing fertilizer usage and eutrophication, which can be done using the soil microbiome. Using potted plants with soils inoculated with *Pseudomonas alcaligenes*, *Pseudomonas denitrificans*, *Bacillus polymyxa*, and *Mycobacterium phlei*, both the shoot and root growth of pea and cotton plants was significantly increased (Egamberdieva & Höflich, 2004). In this study, utilizing a random forest model, the presence or absence of inflorescences of an *Asclepias* (milkweed) plant was predicted using the soil microbiome as an input with 64% accuracy on test data. *Euryarchaeota*, *Acidobacteria*, and *Chlorobi* were identified as the most important phyla in predicting the presence of inflorescences.

Table of Contents

Acknowledgements.....	4
Figure Legend.....	5
Introduction	6
<i>Motivation</i>	6
<i>DNA sequencing – History and Methodology</i>	6
<i>Taxonomic Identification with Metalign</i>	8
<i>The Soil Microbiome and its Effects on Plant Fitness</i>	8
<i>The Soil Microbiome as a Predictor in Machine Learning Models</i>	9
Methods	10
<i>Sample Collection</i>	10
<i>DNA Extraction and Sequencing</i>	10
<i>Taxonomic Identification with Metalign</i>	11
<i>Random Forest Modeling</i>	11
Results	12
<i>Description of Microbiome Data Results</i>	12
<i>Description of Random Forest Modeling results</i>	15
Discussion	21
Conclusion	22
References	23

Acknowledgements

I would like to thank my PIs, Professors Joshua Puzey and Ronald Smith, as well as my committee members Professors William Soto and Dana Willner. I also want to give a profound thank you to the members of Puzey Lab who helped generate the data used for my thesis: Lizzy Davies, Christian D'Orgeix, Natalie Nantais, and many others. Without their work, my thesis would not be possible. I would also like to thank Professor Vasiliu for his help discussing our modeling efforts.

I also want to acknowledge William & Mary Research Computing for providing computational resources and/or technical support that have contributed to the results reported within this paper. URL: <https://www.wm.edu/it/rc>

Figure Legend

Figure 1: Map of sampling locations across Virginia	10
Figure 2: Phylum community composition	12
Figure 3: Species Richness for each taxonomic level comparing zero inflorescences and one or more inflorescences.....	13
Figure 4: Percentage of representative of identified taxa across all samples for each taxonomic level	14
Figure 5: Correlation matrix of all phyla	15
Figure 6: Confusion matrices of Random Forest model predictions for zero and one or more inflorescences	16
Figure 7: Feature importance values of Random Forest model	17
Figure 8: Important features in Random Forest model.	18
Figure 9: Distribution of important features	19
Figure 10: Correlation matrix of important phyla in the random forest model.	20

Introduction

Motivation

The UN estimated that the global population could reach 9.7 billion by 2050 (United Nations). As a result, the amount of food required to feed humanity is thought to double by 2050 (Ray et al., 2012); whereas other studies forecast a 50% to 75% required increase in certain areas of crop production (Prosekov & Ivanova, 2018). From 1985 to 2005, the total amount of crops produced increased by 28%. However, crop yields have begun to stagnate, and even decline, in some parts of the world (Ray et al., 2012). Additionally, climate change threatens current crop production and is expected to cause crop loss in the US, with corn, soybean, and cotton production expected to decrease by 30% to 36% (Schlenker & Roberts, 2009). Cai, Wang, and Laurent (2009) predict the rainfed corn yield in central Illinois to decrease by 23% to 34% by 2055. With the increase in demand, and predicted decline in yields, humanity must find a way to increase crop production.

To keep pace with the current demand for food, the amount of fertilizer used by the United States, India, and China has increased from 9.31 million tons in 1961, to 92.63 million tons in 2006 (Adesemoye & Kloepper, 2009). These numbers are expected to increase as food demand rises. However, fertilizers are not environmentally sustainable methods of boosting crop production, with fertilizer usage resulting in nutrient run-off, causing eutrophication (Adesemoye & Kloepper, 2009). Zhu and Chen (2002) found that up to 19% of Nitrogen in fertilizers is lost to leeching, while 1.5% to 2% was lost to runoff. However, other studies estimate that 82% of fertilizer nutrients are left behind after crop harvests (Khan & Mohammad, 2014).

The primary cause for eutrophication, Phosphorus, has increased in cycling from land to ocean by more than three times, from 8 teragrams to 22 teragrams per year (Howarth & Choi, 2005). Harmful algal blooms, caused by excessive nutrients, can make water unsafe to drink and harm aquatic food sources (Wurtsbaugh et al., 2019). With the increase of demand for food to feed a growing population, novel and sustainable farming methods must be utilized to keep pace with demand while also causing less damage to the surrounding environment. These affected environments also include the soil microbiome, which, to be analyzed, first needs to be sequenced and identified.

DNA sequencing – History and Methodology

DNA sequencing has recently become economical for research usage as the cost of sequencing and time required have drastically decreased while the amount of bases identified and the accuracy have dramatically increased (Mardis, 2017; Shendure et al., 2017). With the advent of Sanger sequencing roughly 40 years ago, development in the field of bioinformatics has moved at a blistering pace to the Next-Generation sequencing (NGS) of today (Shendure et al., 2017).

In 1977, Sanger and his team sequenced the genome of the phiX174 bacteriophage, the first genome to ever be fully sequenced (Men et al., 2008). This genome contained about 5,000 bases of DNA, or 0.005 Mb. In just a couple decades, the human genome was sequenced in

2004 with a size of 3,200 Mb (Lander et al., 2001). The cost of using Sanger Sequencing was roughly \$1 per 1,000 bases, or 1 kb. (Men et al., 2008). From just 2007 to 2012, the cost of sequencing per base decreased by 400% (Shendure et al., 2017). These numerous gains in efficiency were due to the implementation of novel sequencing techniques. However, it is important to start with the past to understand the present.

Sanger Sequencing was the first sequencing methodology, published in 1977 (Men et al., 2008). The DNA sequences were identified by first separating the chains into different lengths by synthesizing new DNA while incorporating ddNTPs to halt any further sequence addition. The ddNTPs are added in a small amount so that the chain termination is random, yielding differentially lengthened sequences. These ends were also capped with radioactive isotopes of phosphorus or sulfur (Shendure et al., 2017). Then, the DNA chains of various lengths were drawn through polyacrylamide gel (PAG) electrophoresis to sort them by size at a resolution up to one base length (Men et al., 2008).

After sorting, the ends were identified using radiography and their positions were read along the four lanes of the gel, one for each base (Shendure et al., 2017). This methodology was time consuming, as the target DNA sequences were not quickly amplified by PCR, but instead were grown as clonal plasmids in bacteria, most commonly *E. coli.*, or as phage vectors (Mardis, 2017). Today, NGS has dramatically evolved output with multiple novel sequencing methods. One of those new methods is Illumina sequencing.

Illumina sequencing has revolutionized sequencing technology by expanding throughput while decreasing costs (Bronner et al., 2013). With Illumina, the DNA is attached to oligonucleotides that protrude from the inside of a glass flow cell. The DNA is then amplified to increase read signals. Next, fluorescently tagged nucleotides are washed over the DNA sequences, each base having a unique color that illuminates when binding to the target DNA sequence. This signal is then read to determine the sequence of the complementary strand. Read lengths can be as long as 250bp per read. The cost to sequence DNA using Illumina sequencing is roughly \$41 per gigabase (gb), or 1,000,000 kb. This technique, sequencing by synthesis, can be used in large industrial sized machines or benchtop machines. The advent of NGS has allowed researchers to conduct research using DNA sequences in different ways to answer novel questions.

The research of microbial life has been revolutionized by NGS (Liu et al., 2021; Nannipieri et al., 2019). Using two flavors of DNA sequencing, amplicon sequencing or shotgun-metagenomics, scientists can study the community compositions of microbiomes in depth (Liu et al., 2021). In amplicon sequencing, the 16S ribosomal DNA is amplified in a sample and sequenced. This method results in fewer reads and a quicker analysis, but at the cost of lower taxonomic resolution.

With metagenomic sequencing, all DNA is just sequenced. This allows for identification down to the species or strain level, while also allowing for analysis beyond simple taxonomic classification and can amplify signals of uncultured bacteria. However, this method is expensive and generates lots more data to analyze than amplicon sequencing. With the advantages of either type of sequencing, scientists can tailor their methodology to better address their research question. After the sequences have been read, they need to identify what organism the DNA came from.

Taxonomic Identification with Metalign

Metalign is a novel taxonomic identification software, with a focus on accuracy and speed (LaPiere, et al., 2020). Metalign efficiently aligns sequences to its extensive 243 GB database, which contains all completed and partial microbial genomes from NCBI, GenBank, and RefSeq. It does not contain any animal or plant genomes. While this extensive genome can result in high accuracies, it would be time consuming to align each sequence to the entire database.

Here, the researchers implement a novel subsetting method called ‘containment min-hash’. This method defines a subset of the database that accurately represents the subset of the sample sequences aligned to the entire database and runs the rest of the samples on the subset database, which is roughly 100 times smaller than the original. For calculating what taxa belong in the subset database, the Jaccard index is calculated and a threshold value of 0.01 or greater is required to be included in the subset data.

Metalign employs Minimap2 as an alignment software, with the strict requirement of 95% perfect sequence alignment to taxonomically identify a sequence. The identified sequences are reported by their relative abundance. Unique mappings count as one point towards the absolute abundance, while multiple mappings to a sequence allocates a proportion of that point compared to the proportion of the unique mappings of each taxa the initial read mapped to. When compared to other identification methods, including Kraken2, DIAMOND, MEGAN6, and others, Metalign consistently produced some of the highest ratios of precision to recall. Metalign was in the middle in terms of required CPU time compared to the other models. Utilizing this software, researchers can explore entire microbiomes and study their compositions across numerous environments.

The Soil Microbiome and its Effects on Plant Fitness

Microbiome communities are not random and are influenced by the host plant and their location relative to it as well as abiotic factors (Bulgarelli et al., 2013; Gopal and Gupta, 2016; Trivedi et al., 2020). Plant microbiome community compositions differ along the rhizosphere and phyllosphere, but are similar between the rhizosphere and bulk soil (Trivedi et al., 2020). Across the rhizosphere and bulk soils, there is a common core community structure.

The largest abundance of bacterial taxa in the rhizosphere and bulk soil are from the phyla Proteobacteria, Actinobacteria, and Acidobacteria, along with many other less abundant phyla (Trivedi et al., 2020). Other microbes have been shown to influence the soil microbiome community composition (Trivedi et al., 2020). These microbes are denoted as ‘hub microorganisms’ and are thought to interact with the plant or other microorganisms to drive changes in community structure.

Soil microbiomes have also been shown to vary based upon abiotic factors, such as pH, soil type (Chang et al., 2017, Jiao et al., 2019), or temperature (Jiao et al., 2019; Luláková et al., 2019). The amount of NH₄ has also been shown to explain the amount of variation in microbial carbon (C_{mic}) by 39.2% and variation in the soil microbiome community composition by 22.8% (McGee et al., 2019). Conversely, through random forest analysis, bacterial diversity has been shown to predict the multi-nutrient cycling index of soil (Jiao et al., 2019).

Different taxa were also identified contributing to specific nutrients, with Euryarchaeota informing the prediction of NO_3^- and pH and Acidobacteria informing the prediction of pH and organic matter. Leveraging these biotic and abiotic interactions, we increase crop production to feed a growing population by tailoring the crop's soil microbiome to promote higher crop yields and increasing crop fitness.

Plant microbiomes have been shown to increase plant health, by contributing to disease suppression (Peralta et al., 2018; Wei et al., 2019), growth and development (Panke-Buisse et al., 2016), and nutrient uptake (Taffner et al., 2020; Trivedi et al., 2020). Plant growth promoting rhizobacteria have been shown to increase many facets of plant health, with the increased uptake of nutrients thought to arise from stimulation of root formation leading to bigger root systems with more root hairs (Adesemoye & Kloepper, 2009). Arbuscular mycorrhiza fungi (AMF) have been shown to increase phosphorus uptake in plants. By engineering soils comprising these beneficial microbes, researchers can create crops with higher yields.

Using potted plants with soils inoculated with *Pseudomonas alcaligenes*, *Pseudomonas denitrificans*, *Bacillus polymyxa*, and *Mycobacterium phlei*, both the shoot and root growth of pea and cotton plants was significantly increased (Egamberdieva & Höflich, 2004). These artificial microbiomes, curated by the researchers, were shown to increase plant health. However, *in vivo*, these communities are dynamic and affected by many factors, including the host plant. With the vast array of microbial taxa affecting the plant directly, indirectly, and through combinations of other interactions, it is important to disentangle and identify the microorganisms contributing to these beneficial interactions. This identification can be made with machine learning models.

The Soil Microbiome as a Predictor in Machine Learning Models

Researchers have utilized machine learning models with microbiomes as the input features to predict important host health traits (Chang et al., 2017; Dong et al., 2020; Weinroth et al., 2019). Weinroth et al. (2019) investigated the formation of liver abscesses in cattle treated with tylosin while comparing their soil and fecal microbiomes. The fecal microbiomes did not differ, but the soil microbiome around the cows did. Utilizing a Least Absolute Shrinkage and Selection Operator (LASSO) model, the researchers chose six microbial taxa to evaluate for correlations to liver abscess: *Euryarchaeota*, *Fibrobacteres*, *candidate phyla Cloacimonetes* [WWE1], WPS2, *Deferribacteres* and *Firmicutes*.

These identified microbial communities at the phylum level in both the fecal and soil samples explained 75% of the variation in presence of liver abscesses. LASSO regression models have also been used to predict disease in humans (Dong et al., 2020). Using the gut microbiome, along with the age and sex of the patient, the LASSO model created by Dong et al. (2020) was able to predict with 80% accuracy if a person had Parkinson's disease.

Other machine learning methods have also been applied to agriculture, with Chang et al (2017) utilizing a Random Forest model to predict whether an area in a crop field was a high or low productive plot by examining the soil microbiome. After investigating the soil characteristics and using a logistic regression that showed no association with crop productivity and 26 tested soil characteristics, the researchers ran a PCA that showed a difference in the soil

microbiome community composition between high and low productive fields. The Random Forest model, evaluating the soil microbiome at the taxonomic level of Order, had an accuracy of 79% when predicting field productivity. The model also identified multiple nitrogen utility-related taxa as important to making the prediction, such as the phyla Actinobacteria, Proteobacteria, and Cyanobacteria.

These studies have shown that machine learning can be used to identify potential interactions between the soil microbiome and the host. These interactions can then be further explored to determine their effects on the host plant and what artificially created microbiome will create healthier plants. **In this study, I will use the soil microbiome as the features in a machine learning model to predict whether an *Asclepias* plant has zero or 1 or more inflorescences.**

Methods

Sample Collection

Samples of *Asclepias* (Common Milkweed) were gathered in the summers of 2020 and 2021 in locations throughout Virginia (Figure 1).

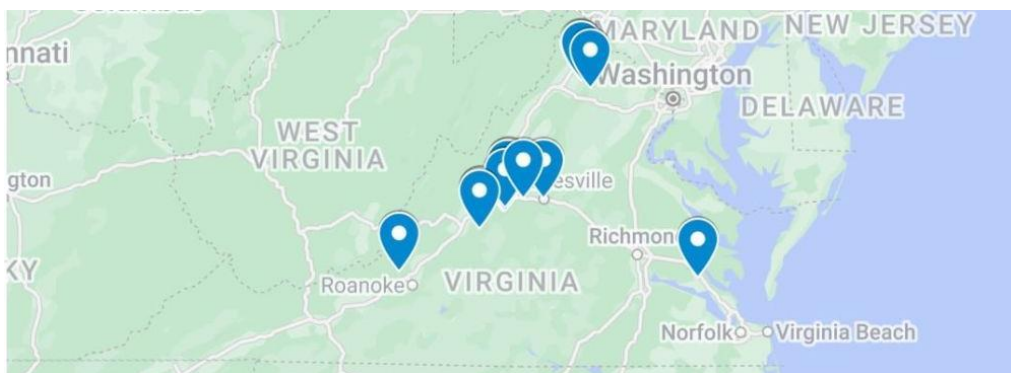


Figure 1: Map of sampling locations across Virginia. In total, 14 sites were used in this analysis. Sites were sampled over the summer months in 2020 and 2021.

At each site, the *Asclepias* plants were measured for phenotypic traits, such as height and number of leaves, and soil cores were taken near the plant, but far enough to minimize root mass being collected in the soil core. The leaf and soil samples were then frozen and brought to the lab for DNA extraction. Soil samples were sent to Waypoint Analytics to determine elemental composition of the soil and other factors such as organic matter content.

DNA Extraction and Sequencing

DNA extraction was carried out using the FastDNA Spin Kit for Soil. The DNA was then cleaned using Zymo Research DNA clean and concentrator because the initial DNA libraries sent were not easy to sequence and the cleaning kit was used to remove expected small, charged molecules from the DNA. The DNA was then packed into libraries and sent to Michigan State

University and sequenced on a Novaseq 6000 S4 lane with paired-end 150 bp reads (University of Oregon GC3F facility).

Taxonomic Identification with Metalign

The quality scores of the received paired-end fastq files were examined to ensure clean sequencing. Once the files were evaluated, the paired-end sequence files were concatenated. The concatenated files were then uploaded to the William & Mary High Performance Cluster to run the identification software Metalign on each file using the mode “sensitive”. Metalign has been shown to classify taxa accurately and quickly (LaPierre et al., 2020).

Random Forest Modeling

After the sequences were run through Metalign, the data were input into a Random Forest model for predicting whether any inflorescences were present on the *Asclepias* sample based upon the input microbiome. A grid search was performed to identify the best out of this list of hyperparameters: taxonomic level, number of trees and max tree depth. The taxa in the dataset were also described by their percent presence across all samples and this became another hyperparameter labeled percent required. A grid search was conducted from 0% to 70% required identification across all samples in 10% increments.

The best performing model constituted 300 trees, with a max depth of 1, and all other hyperparameters were set to default. There was no increase in accuracy across the taxonomic levels, so phylum was used to shorten runtime. The optimum percent presence value was 0%. Therefore, the input data consisted of all 142 phyla identified, regardless of their rarity across all samples. Due to a class imbalance of 75% of the samples having at least one inflorescence, the parameter “class_weight” was set to “balanced_subsample”. This parameter will add weights to the Gini Impurity calculated in the tree based upon the composition of classes in the bootstrapped subset. No scaling or transformations were applied to the data. The model was then put through Leave One Out cross-validation and train and test scores were calculated as averages across all k-folds.

Results

Description of Microbiome Data Results

The dataset consisted of 156 plant samples with their number of inflorescences quantified and soil cores extracted. A large proportion of the plants had one or more inflorescences, 75%, or 116 samples, which created an imbalanced dataset with the remaining 25%, or 40 samples, having no inflorescences. Soil characteristics and plant measurements were found to differ significantly by site, with all one-way ANOVAs resulting in $p < 0.05$. Collectively, 142 phyla, 164 classes, 372 orders, 836 families, 3,017 genera, and 22,645 species were identified across all soil samples using Metalign. The strain identifications were reported by Metalign; however, only the most accurately identified strain is reported. Therefore, the number of strains identified is artificially low as other strains of the same species identified at high accuracies would not be reported. The community composition when examining phyla across all samples was dominated by *Proteobacteria* and *Actinobacteria* (Figure 2).



Figure 2: Phylum community composition. This figure contains two samples from each site and plots the relative abundance of each taxon. The communities were dominated in abundance by *Proteobacteria* and *Actinobacteria*. The communities also displayed similar compositions throughout all samples except for three, which are not depicted here.

The samples also displayed a similar distribution in species richness between zero inflorescences and one or more inflorescences plants across all taxonomic levels (Figure 3).

Species richness for each taxonomic level

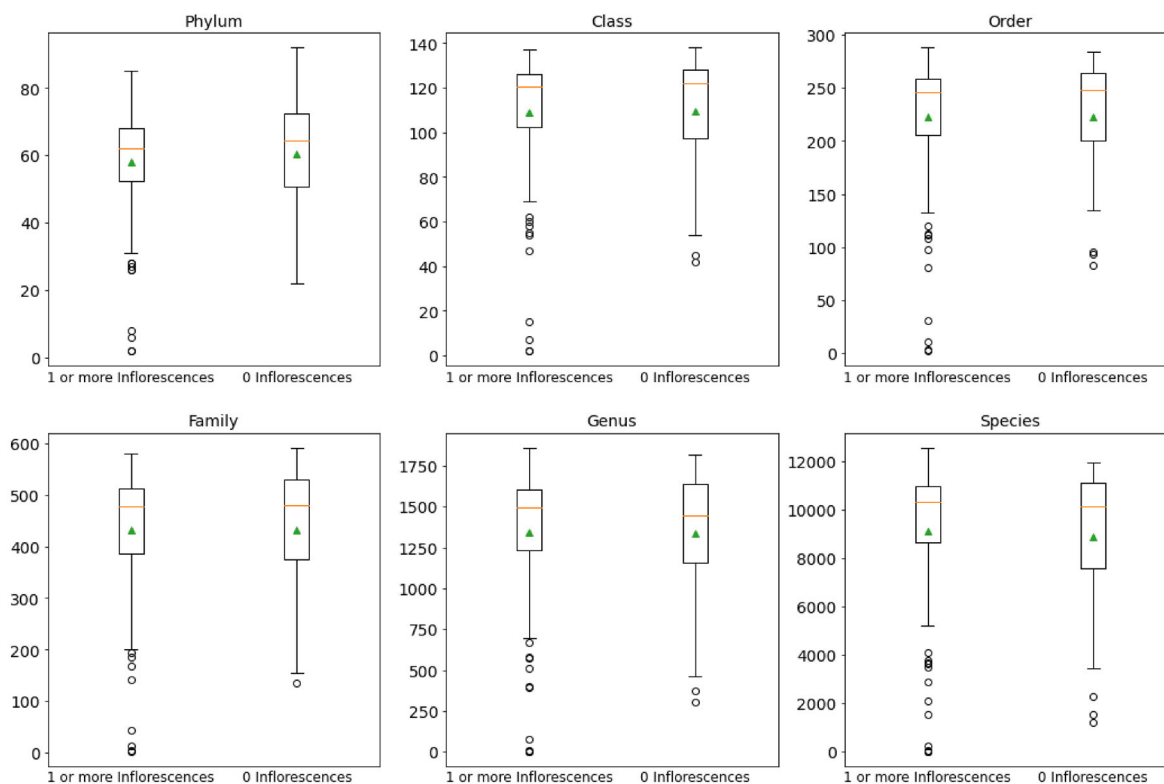


Figure 3: Species Richness for each taxonomic level comparing zero inflorescences and one or more inflorescences. The distribution of species richness across samples is similar between zero and one or more inflorescences in all taxonomic levels.

The level of phylum also displayed a high percentage of rarer taxa compared to the other taxonomic levels (Figure 4).

Percent representation of identified taxa across all samples for each taxonomic level

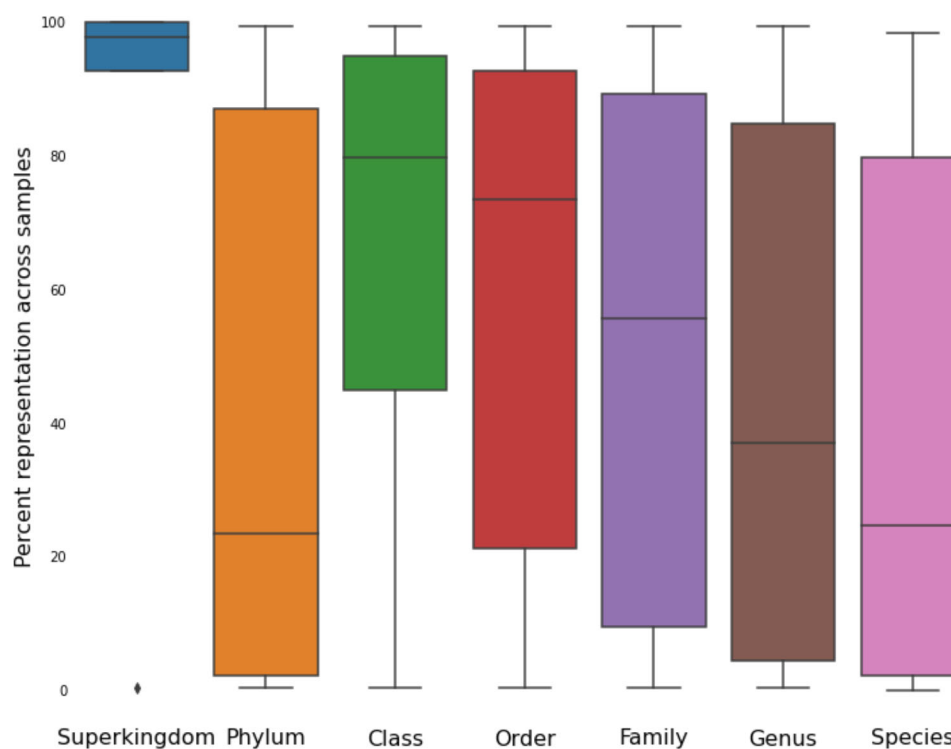


Figure 4: Percentage of representative of identified taxa across all samples for each taxonomic level. The level of phylum displayed the largest variation in percent representation for each taxon across all samples and also contains the lowest average of around 22% average representation across all samples. One sample contained one identification of a viroid.

The microbiome displayed high levels of multicollinearity, which could complicate analysis of important taxa (Figure 5) as correlation vs causation cannot be discerned by feature importance in the model.

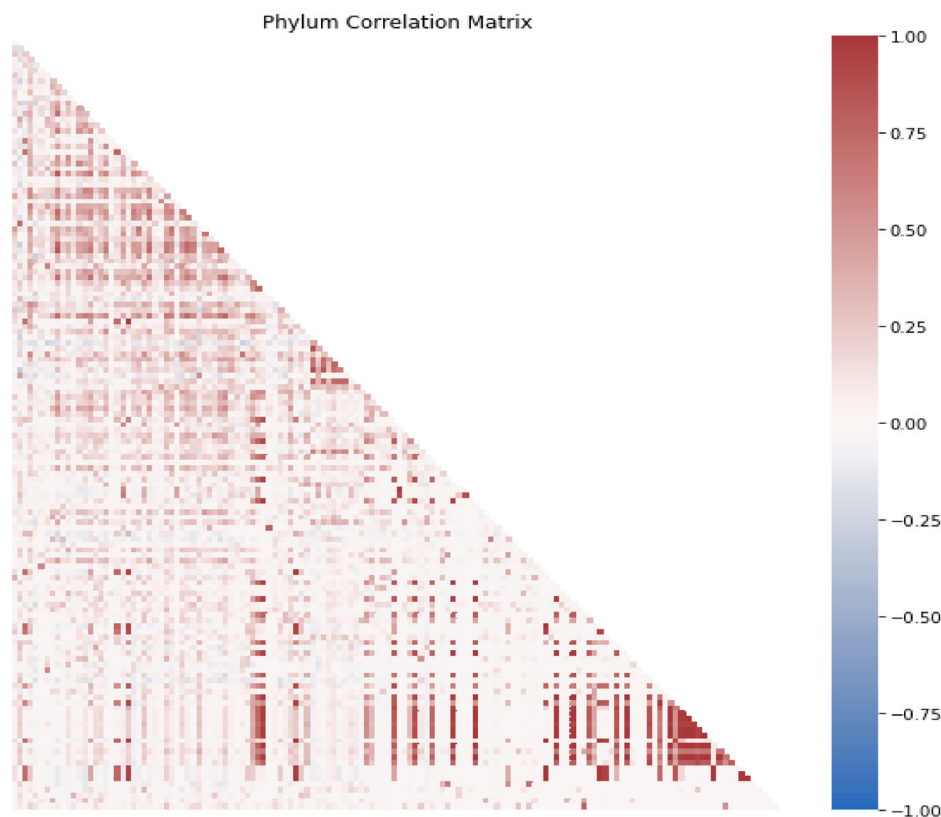


Figure 5: Correlation matrix of all phyla. Overall, some microbe groups were highly correlated with others while there also being numerous smaller correlations throughout. Interestingly, there were not many if any negative correlations seen.

Description of Random Forest Modeling results

The resulting Random Forest Classifier model was put through leave-one-out cross validation on 156 observations, which contained relative abundance values of 142 phyla identified across the samples. To identify the optimal model parameters, a grid search was conducted across the following model and data parameters: number of trees, model depth, required percent presence across samples, and taxonomic level. The model parameters used in the final model consisted of 300 trees, a max depth of 1, zero taxa dropped from the data frame, and using the taxonomic level of phylum. The target variable, originally ranging from zero to ten, was changed to a binary variable where zero was all values equal to zero, and one if the values were above zero. This equates to predicting the presence of inflorescences. Through cross validation, the model had a training accuracy of 0.78% and testing accuracy of 0.64%. A confusion matrix was produced to describe the effectiveness of the model at predicting certain classes by their counts as well as by the percentage of that class overall (Figure 6).

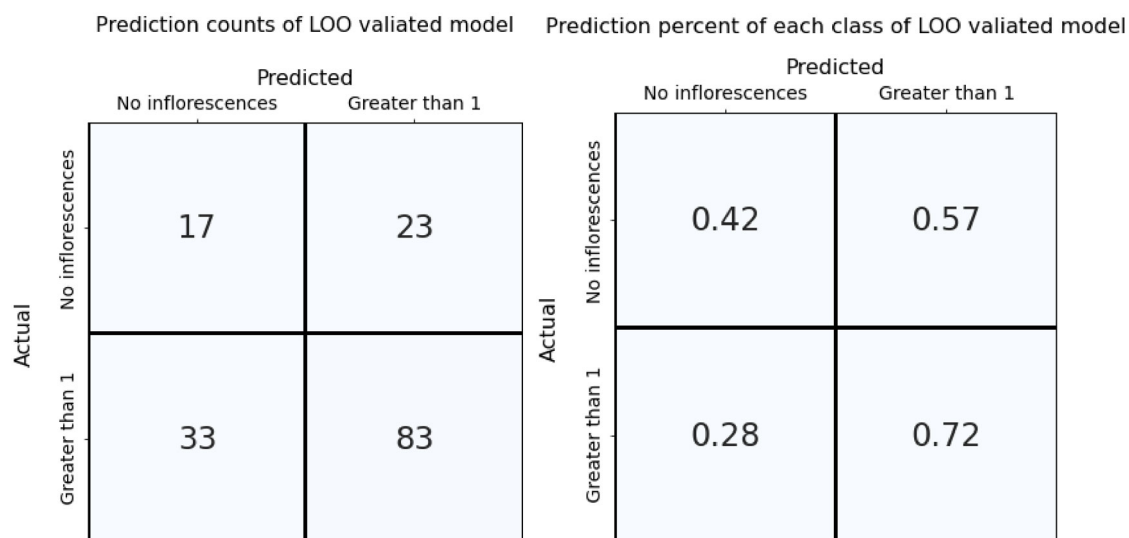


Figure 6: Confusion matrices of Random Forest model predictions for zero and one or more inflorescences. The matrix on the left details the count of class predictions by the Random Forest model. The model produced more false positive predictions than true negatives. The model was more accurate when predicting true positives overall, but still made many false negative predictions. These predictions are also reflected in the left matrix, which displayed the prediction counts as a percentage of the actual class.

Next, the important features of the model were determined by averaging the importance values across all 300 trees in the model for each fold in the leave-one-out cross-validation. Then, the importance values for each fold were also averaged. The result showed a clear top three phyla: *Euryarchaeota*, *Acidobacteria*, and *Chlorobi*, and 98 of 142 phyla had non-zero importance values (Figure 7).

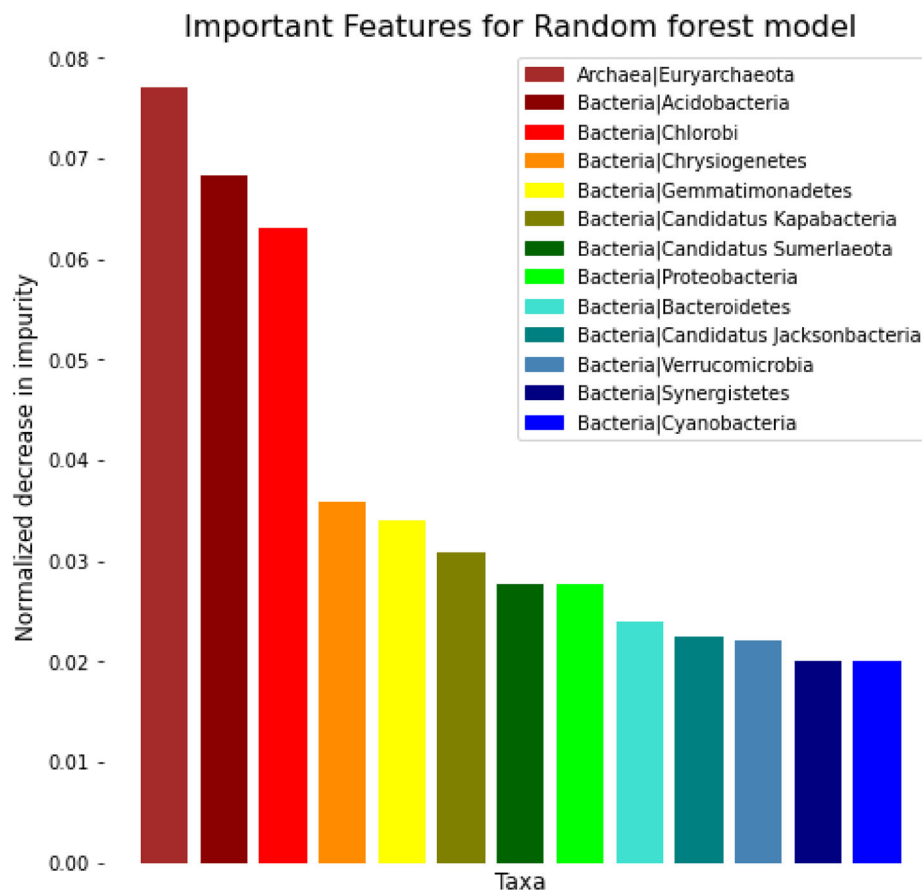


Figure 7: Feature importance values of Random Forest model. The top 13 most important features from the model ran show a clear top three most important phyla when predicting the presence of inflorescences.

Due to variance in the ranking of important features when running the models using the same hyperparameters and different random states, the feature importances of 50 models through leave-one-out cross-fold validation were calculated to elucidate more clearly consistently important features. The importance features across all 300 trees were averaged for 50 models run through leave-one-out cross-fold validation and then those averages were averaged to determine the consistently important input features for the model. Of the 142 phyla in the dataset, 130 had an importance value greater than zero. The results for the top 13 informative taxa from the 50 models ran are reported in figure 8. The distribution of those importance values across all 142 phyla is reported in figure 9.

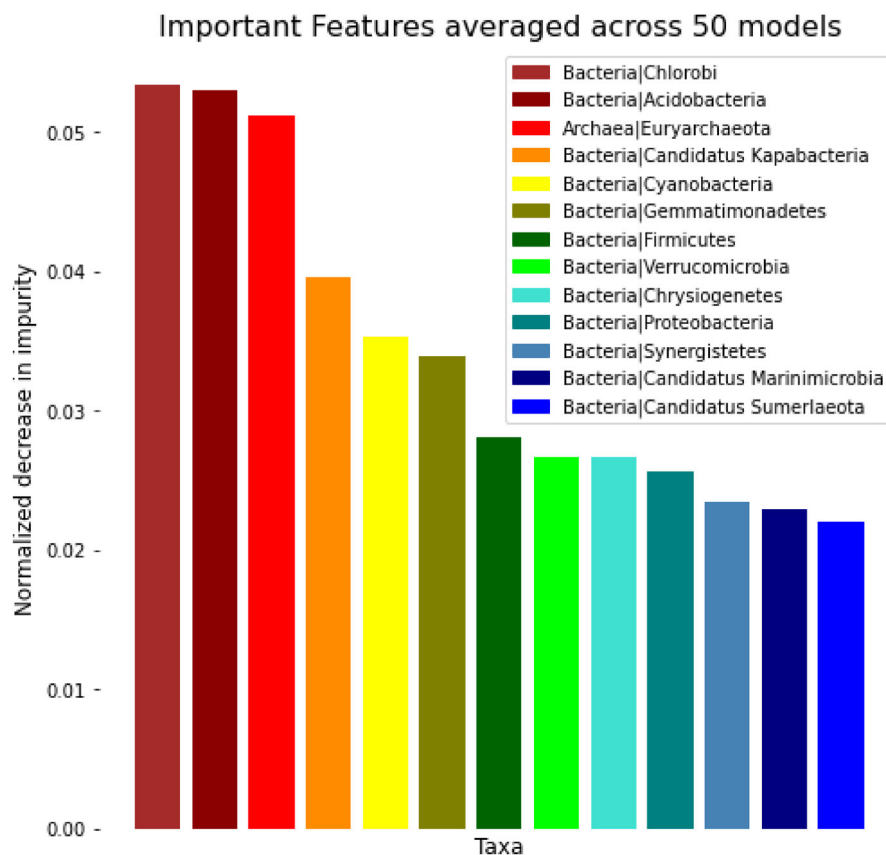


Figure 8: Important features in Random Forest model. These are the top 13 features for predicting the presence of inflorescences averaged across 50 models. All values had a low level of importance, with the top three features slightly standing out from the rest.

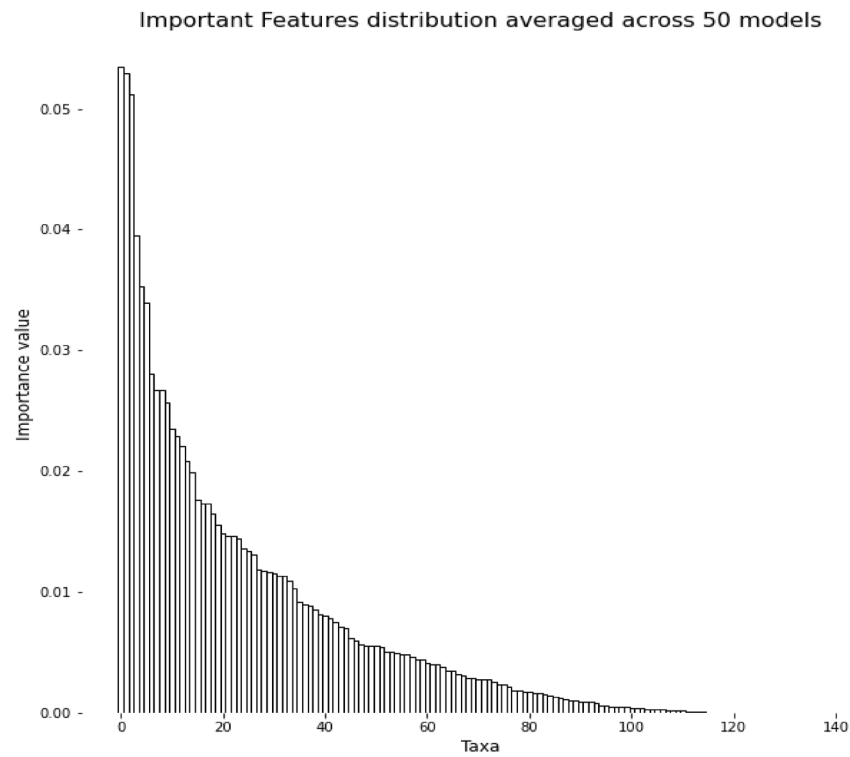


Figure 9: Distribution of important features. The important features show the same top three and then a descent down in importance value.

The abundance of the 13 most important phyla also exhibited strong correlations with one another (Figure 10).

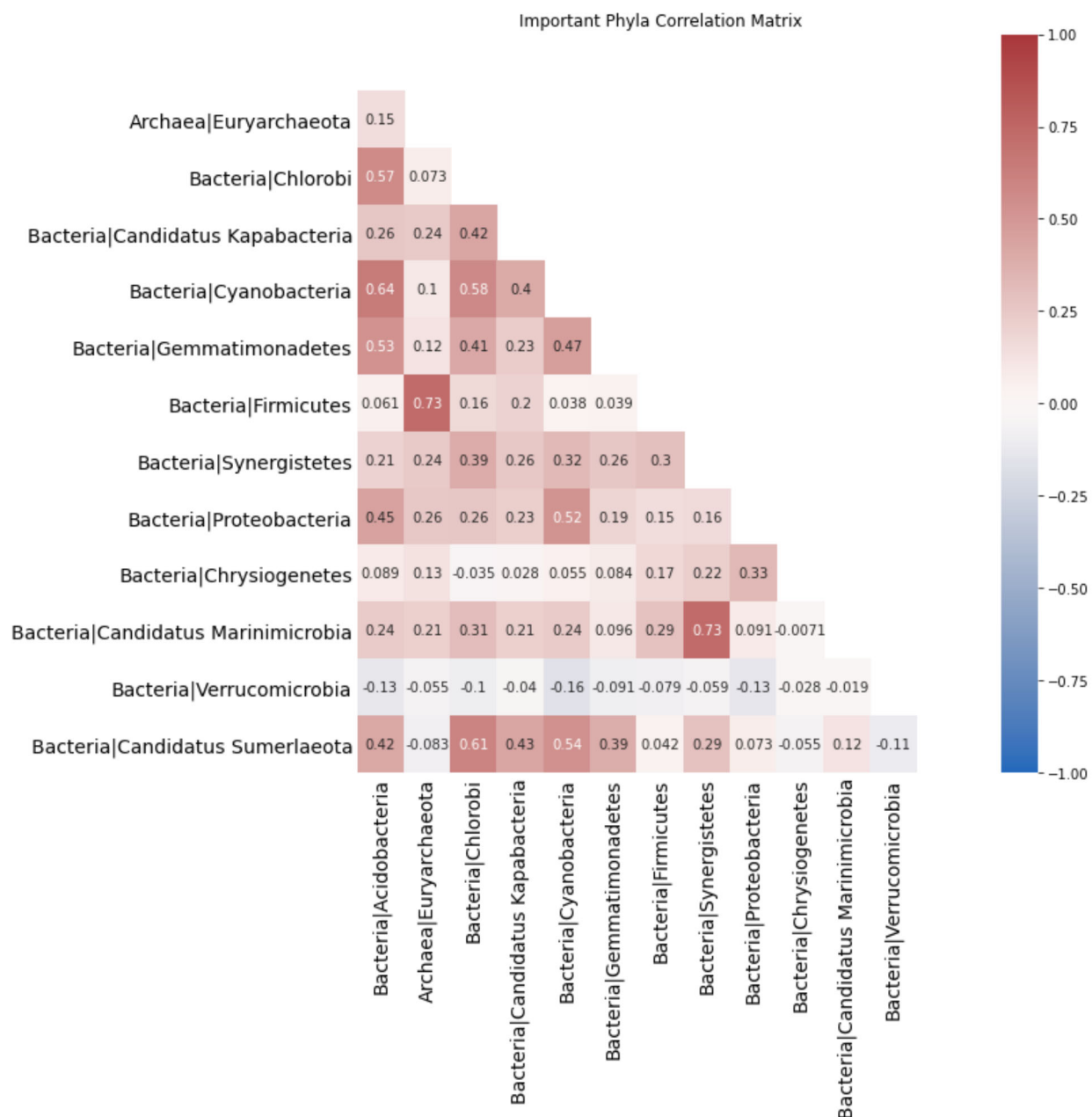


Figure 10: Correlation matrix of important phyla in the random forest model. Some phyla showed strong correlations with numerous other phyla, such as Acidobacteria.

Discussion

As discussed previously, soil microbiomes are reactive communities that respond to biotic and abiotic factors (Jiao et al., 2019; Luláková et al., 2019). This can present an issue when collecting soil samples for analysis, as the dynamic nature of the microbiome could result in different conditions being sampled across different sites. Sampling at numerous timepoints could account for such dynamism (Nannipieri et al., 2019).

A second consideration when sampling is the resilience of the different taxa to change. Bacterial abundances have been shown to vary in their resilience to environmental effects based upon their abundance, with taxa in large abundance resilient to change and rare taxa more sensitive (Jiao et al., 2019). The study also showed higher levels of resilience in communities with higher levels of diversity. If sampling at a site occurred right after a disturbance, then the rarer taxa could be missed as they were more susceptible to the change, resulting in lower community diversity in the sequenced microbial communities.

Along with the core communities around different locations of the plant, there are also predictable differences in microbiome diversity in the vertical distribution of microbe communities. Jiao et al. (2018) reported that as depth increases in the soil from 0 to 80 cm the bacterial and fungal diversity decreases, while the archaeal diversity increases. Bacterial beta-diversity was linked to multi-nutrient cycling in the deep soil layers and the archaeal beta-diversity was linked to the superficial soil layers. Therefore, the depth of the soil sample taken can influence the diversity of the sampled microbiome. Compared with the different depths of plant roots, standardized sampling practices need to be made to ensure the same area of the microbiome is captured. The bulk soil microbiome has also been shown to have a higher alpha diversity than the rhizospheres (Wu et al., 2018). Therefore, while collecting soil cores the researcher has to balance not gathering much root material while also minimizing the distance away from the roots to capture more of the rhizosphere than the bulk soil microbiome.

Acidobacteria was one of the most important phyla in informing the model prediction. This phylum has been shown to increase in abundance as plant growth increases, but to also increase in abundance in the increasing abundance of PGPR (Kalam et al., 2017). *Acidobacteria* abundance was correlated with nearly all other important taxa (*Euryarchaeota*: 0.15 and *Chlorobi*: 0.57). *Acidobacteria* abundance has also been shown to correlate with stages of plant development (Chaparro, Badri, & Vivanco, 2014). *Acidobacteria* abundance increased from plant seed to vegetative developmental state, but then decreased once the plant started bolting and flowering.

Euryarchaeota was identified as another important phylum in the model. *Euryarchaeota* is a highly diverse phylum of Archaea that has taxa utilizing many different metabolic pathways (Bomberg & Timonen, 2007). *Euryarchaeota* can be found in the rhizosphere and are not uncommonly the most abundant phylum. This taxon is thought to be correlated with the presence of mycorrhizal fungi and grows in older root tips over growing roots. When evaluating the correlations within the sampled microbiome, *Euryarchaeota* had a correlation of 0.48 to the mycorrhizal fungi phylum *Mucoromycota*. Archaea have been shown through metagenomic analysis to influence the host plant in three ways: nutrient supply for the plant, promotion of plant growth via biosynthesis of auxin, and competition and syntrophic interactions with fungi and bacteria (Taffner et al., 2020).

The third most important phylum was *Chlorobi*. This phylum consists of green sulfur bacteria, most of which are phototrophic anaerobes and found in low-light environments (Bryant et al., 2012). Most members of this phylum can grow using only N₂ as a source of Nitrogen and all require sulfide for their metabolism. This phylum has been identified as an endophyte (Kuffner et al., 2010); however, there is not much literature on its effects on plant growth. This phylum was possibly selected as important to model predictions because it is highly correlated with *Acidobacteria* (0.57).

For future research, other plants characteristics may be predicted based upon the soil microbiome. The fitness of the plant may also be quantified based upon the plant's characteristics and that value could be predicted as an overall view of how the microbiome affects the plant's health. In the future, the researchers may switch to using regression models such as LASSO or ElasticNet. These models can remove features that do not contribute information to the prediction of the target, which is useful with datasets like in this study that have a plethora of features. Correlation terms would also be interesting to explore as certain microbes have been shown to interact with plants only in the presence of another specific taxa.

Conclusion

As shown before, machine learning algorithms can detect patterns between microbiome community structures and host phenotypes, while also identifying important microbes within those communities as contributing the most predictive information. The important phyla for informing the model prediction have been shown to influence plant health and they have been shown to be affected by plant developmental stages (Chaparro, Badri, & Vivanco, 2014; Kalam et al., 2017; Taffner et al., 2020). In this study, the resulting soil microbiomes showed community compositions expected for the rhizosphere, with *Proteobacteria* and *Actinobacteria* dominating in abundance. Using a Random Forest model, the microbiome was able to contribute information in predicting the number of inflorescences. In the future, regression models might be used to yield better interpretations of taxa importance on model predictions. The important concept that microbes can predict host phenotypes underscores the importance of researching host microbiome relationships. In the case for agriculture, the soil microbiome can result in increased crop production by an environmentally sustainable method.

References

- Adesemoye, A. O., & Kloepper, J. W. (2009). Plant–microbes interactions in enhanced fertilizer-use efficiency. *Applied Microbiology and Biotechnology*, 85(1), 1–12. <https://doi.org/10.1007/s00253-009-2196-0>
- Bomberg, M., & Timonen, S. (2007). Distribution of Cren-and Euryarchaeota in Scots pine mycorrhizospheres and boreal forest humus. *Microbial ecology*, 54(3), 406–416.
- Bronner, I. F., Quail, M. A., Turner, D. J., & Swerdlow, H. (2013). Improved protocols for illumina sequencing. *Current protocols in human genetics*, 79(1), 18–2.
- Bryant, D. A., Liu, Z., Li, T., Zhao, F., Costas, A. M. G., Klatt, C. G., ... & Overmann, J. (2012). Comparative and functional genomics of anoxygenic green bacteria from the taxa Chlorobi, Chloroflexi, and Acidobacteria. In *Functional genomics and evolution of photosynthetic systems* (pp. 47–102). Springer, Dordrecht.
- Bulgarelli, D., Schlaeppi, K., Spaepen, S., van Themaat, E. V., & Schulze-Lefert, P. (2013). Structure and functions of the bacterial microbiota of plants. *Annual Review of Plant Biology*, 64(1), 807–838. <https://doi.org/10.1146/annurev-arplant-050312-120106>
- Cai, X., Wang, D., & Laurent, R. (2009). Impact of climate change on crop yield: A case study of rainfed corn in central Illinois. *Journal of Applied Meteorology and Climatology*, 48(9), 1868–1881.
- Chang, H.-X., Haudenschild, J. S., Bowen, C. R., & Hartman, G. L. (2017). Metagenome-wide association study and machine learning prediction of bulk soil microbiome and crop productivity. *Frontiers in Microbiology*, 8. <https://doi.org/10.3389/fmicb.2017.00519>
- Chaparro, J. M., Badri, D. V., & Vivanco, J. M. (2014). Rhizosphere microbiome assemblage is affected by plant development. *The ISME journal*, 8(4), 790–803.
- Dong, M., Li, L., Chen, M., Kusalik, A., & Xu, W. (2020). Predictive analysis methods for human microbiome data with application to parkinson’s disease. *PLOS ONE*, 15(8). <https://doi.org/10.1371/journal.pone.0237779>
- Egamberdiyeva, D., & Höflich, G. (2004). Effect of plant growth-promoting bacteria on growth and nutrient uptake of cotton and pea in a semi-arid region of Uzbekistan. *Journal of Arid Environments*, 56(2), 293–301.
- Gopal, M., & Gupta, A. (2016). Microbiome selection could spur next-generation plant breeding strategies. *Frontiers in Microbiology*, 7. <https://doi.org/10.3389/fmicb.2016.01971>

- Howarth, R., & Choi, E. (2005). Nutrient Management. *Ecosystems and Human Well-Being: Policy Responses: Findings of the Responses Working Group*, 3, 295.
- Jiao, S., Chen, W., Wang, J., Du, N., Li, Q., & Wei, G. (2018). Soil microbiomes with distinct assemblies through vertical soil profiles drive the cycling of multiple nutrients in reforested ecosystems. *Microbiome*, 6(1). <https://doi.org/10.1186/s40168-018-0526-0>
- Jiao, S., Wang, J., Wei, G., Chen, W., & Lu, Y. (2019). Dominant role of abundant rather than rare bacterial taxa in maintaining agro-soil microbiomes under environmental disturbances. *Chemosphere*, 235, 248–259. <https://doi.org/10.1016/j.chemosphere.2019.06.174>
- Kalam, S., Das, S. N., Basu, A., & Podile, A. R. (2017). Population densities of indigenous Acidobacteria change in the presence of plant growth promoting rhizobacteria (PGPR) in rhizosphere. *Journal of basic microbiology*, 57(5), 376-385.
- Khan, M. N., & Mohammad, F. (2014). Eutrophication: challenges and solutions. In *Eutrophication: causes, consequences and control* (pp. 1-15). Springer, Dordrecht.
- Kuffner, M., De Maria, S., Puschenreiter, M., Fallmann, K., Wieshammer, G., Gorfer, M., ... & Sessitsch, A. (2010). Culturable bacteria from Zn-and Cd-accumulating *Salix caprea* with differential effects on plant growth and heavy metal availability. *Journal of applied microbiology*, 108(4), 1471-1484.
- Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., ... & Proctor, M. J. (2001). Initial sequencing and analysis of the human genome.
- LaPierre, N., Alser, M., Eskin, E., Koslicki, D., & Mangul, S. (2020). Metalign: efficient alignment-based metagenomic profiling via containment min hash. *Genome biology*, 21(1), 1-15.
- Liu, Y. X., Qin, Y., Chen, T., Lu, M., Qian, X., Guo, X., & Bai, Y. (2021). A practical guide to amplicon and metagenomic analysis of microbiome data. *Protein & cell*, 12(5), 315-330.
- Luláková, P., Perez-Mon, C., Šantrůčková, H., Ruethi, J., & Frey, B. (2019). High-alpine permafrost and active-layer soil microbiomes differ in their response to elevated temperatures. *Frontiers in Microbiology*, 10. <https://doi.org/10.3389/fmicb.2019.00668>
- Mardis, E. R. (2017). DNA sequencing technologies: 2006–2016. *Nature protocols*, 12(2), 213-218.
- McGee, K. M., Eaton, W. D., Porter, T. M., Shokralla, S., & Hajibabaei, M. (2019). Soil microbiomes associated with two dominant Costa Rican tree species, and

- implications for remediation: A case study from a Costa Rican conservation area. *Applied Soil Ecology*, 137, 139–153. <https://doi.org/10.1016/j.apsoil.2019.02.007>
- Men, A. E., Wilson, P., Siemering, K., & Forrest, S. (2008). Sanger DNA sequencing. *Next Generation Genome Sequencing: Towards Personalized Medicine*, 1-11.
- Nannipieri, P., Penton, C. R., Purahong, W., Schloter, M., & van Elsas, J. D. (2019). Recommendations for soil microbiome analyses. *Biology and Fertility of Soils*, 55(8), 765–766. <https://doi.org/10.1007/s00374-019-01409-z>
- Panke-Buisse, K., Lee, S., & Kao-Kniffin, J. (2016). Cultivated sub-populations of soil microbiomes retain early flowering plant trait. *Microbial Ecology*, 73(2), 394–403. <https://doi.org/10.1007/s00248-016-0846-1>
- Peralta, A. L., Sun, Y., McDaniel, M. D., & Lennon, J. T. (2018). Crop rotational diversity increases disease suppressive capacity of soil microbiomes. *Ecosphere*, 9(5). <https://doi.org/10.1002/ecs2.2235>
- Prosekov, A. Y., & Ivanova, S. A. (2018). Food security: The challenge of the present. *Geoforum*, 91, 73–77. <https://doi.org/10.1016/j.geoforum.2018.02.030>
- Ray, D. K., Ramankutty, N., Mueller, N. D., West, P. C., & Foley, J. A. (2012). Recent patterns of crop yield growth and stagnation. *Nature communications*, 3(1), 1-7.
- Schlenker, W., & Roberts, M. J. (2009). Nonlinear temperature effects indicate severe damages to US crop yields under climate change. *Proceedings of the National Academy of sciences*, 106(37), 15594-15598.
- Shendure, J., Balasubramanian, S., Church, G. M., Gilbert, W., Rogers, J., Schloss, J. A., & Waterston, R. H. (2017). DNA sequencing at 40: past, present and future. *Nature*, 550(7676), 345-353.
- Taffner, J., Bergna, A., Cernava, T., & Berg, G. (2020). Tomato-associated archaea show a cultivar-specific rhizosphere effect but an unspecific transmission by seeds. *Phytobiomes Journal*, 4(2), 133-141.
- Trivedi, P., Leach, J. E., Tringe, S. G., Sa, T., & Singh, B. K. (2020). Plant–Microbiome Interactions: From Community Assembly to Plant Health. *Nature Reviews Microbiology*, 18(11), 607–621. <https://doi.org/10.1038/s41579-020-0412-1>
- United Nations, Department of Economic and Social Affairs, Population Division (2019). *World Population Prospects 2019: Highlights*. ST/ESA/SER.A/423
- Wurtsbaugh, W. A., Paerl, H. W., & Dodds, W. K. (2019). Nutrients, eutrophication and harmful algal blooms along the freshwater to marine continuum. *Wiley Interdisciplinary Reviews: Water*, 6(5), e1373.

- Wei, Z., Gu, Y., Friman, V.-P., Kowalchuk, G. A., Xu, Y., Shen, Q., & Jousset, A. (2019). Initial soil microbiome composition and functioning Predetermine Future Plant Health. *Science Advances*, 5(9). <https://doi.org/10.1126/sciadv.aaw0759>
- Weinroth, M. D., Martin, J. N., Doster, E., Geornaras, I., Parker, J. K., Carlson, C. R., Metcalf, J. L., Morley, P. S., & Belk, K. E. (2019). Investigation of tylosin in feed of feedlot cattle and effects on liver abscess prevalence, and fecal and soil microbiomes and Resistomes. *Journal of Animal Science*, 97(11), 4567–4578. <https://doi.org/10.1093/jas/skz306>
- Wu, S.-H., Huang, B.-H., Huang, C.-L., Li, G., & Liao, P.-C. (2017). The aboveground vegetation type and underground soil property mediate the divergence of soil microbiomes and the biological interactions. *Microbial Ecology*, 75(2), 434–446. <https://doi.org/10.1007/s00248-017-1050-7>
- Zhu, Z. L., & Chen, D. L. (2002). Nitrogen fertilizer use in China—Contributions to food production, impacts on the environment and best management strategies. *Nutrient Cycling in Agroecosystems*, 63(2), 117-127.