**Main points of the paper:**

* Why?
  + Current research low throughput and expensive
  + 16S rRNA and 18S rRNA the most important of the genes we chose
* What have others done?
* What did we do?
  + Elastic net based on prev paper results
* Main results
  + Validation model metrics
    - The metrics of the validation models for each method shouldn’t be taken at face value because removing non-overfit features from consideration will remove some information and increase the error anyway. What’s important is that the validation models’ metrics were in the same ballpark.
    - No waveband selection method gave the best (or worst) validation metrics for all five genes.
    - The validation models could be used of course, but considering the relatively lightweight elastic net models involved, might as well just use the whole dataset.
    - 16S consistently predictable with good metrics. 18S consistently not predictable with this data and methodology; all models worse than blindly guessing the mean every time. The other three genes somewhat predictable, but not super good.
    - Restricting input to visible light gave better results than using entire dataset for 16S (and ureC). Didn’t increase the error by much for the other three either. This is the biggest practical positive result of this study.
  + Waveband selections
    - Even though the hyperspectral data had a spectral resolution of 1 nm, waveband selection at this level of precision has questionable generalizability. Wavebands tended to have very large Pearson correlations with neighboring wavebands for very large regions of the spectrum, particularly in the near infrared (NIR) region <cross reference to corr figure>. At this level, any variation in waveband selections might be due more to noise or stochastic algorithm choices than a true signal in the data. With correlation greater than 0.999 for adjacent wavebands across the spectrum, such wavebands become practically interchangeable. For this reason, we report waveband selection results at a coarser resolution, rounding to the nearest 10 nm.
    - The regions in <cross reference to corr> with higher correlations tended to be regions with high reflectance, and vice versa.
    - Generally no strong agreement for important wavebands among methods and genes, but a few patterns emerged
    - 16S (most important along with 18S, but also best model results). Looking at consensus and baseline models in particular. Order results by decreasing coefficients
      * Visible light only:
        + Consensus: one 620 (orange), then four at 590-600 (orange), then three at 530-540 (green).
        + Baseline: 440 (adjacent extreme coefficients) (purple), 540 (green), 590-600 (orange)
        + The pattern looks basically the same for both consensus and baseline, from the green light onward (ignoring purple/blue). This strengthens results
      * All wvs: 1890-1940 (FSWIR) particularly important
    - For genes overall: 510-560 (green), 700-720 (RE), 970-980 (NIR), 1890-1940 (FSWIR; this from 16S) notable clusters for all.
    - For genes overall: <TODO: point out clusters for consensus, vis-only, all genes>
* Limitations
  + Small data
  + Controlled conditions; risk of poor generalization to actual conditions
  + Very high correlations among adjacent wavebands (which makes sense) – but we did address this problem throughout, so maybe this shouldn’t be put here
  + Somewhat arbitrary numbers of wavebands to be selected
* Future work
  + Generalize models to coarser spectral resolution for compatibility with more affordable equipment. This could possibly be extended to broadband (is this the right term?) multispectral or visible light cameras for more practical applications
  + Other genes
  + Other plants
  + Field study instead of greenhouse
  + More data to build confidence in model robustness/generalization

**Abstract**

<abstract here>

**Introduction**

Having a balanced soil microbiome is critical for growing healthy crops. However, monitoring the microbiome traditionally requires laborious data collection with expensive, low-throughput laboratory analysis. Recently, a prototype method was developed to monitor the soil microbiome via hyperspectral measurements of plants growing in that soil <brooks2024>. Several machine learning models using hyperspectral reflectance of cotton (*Gossypium hirsutum*) leaves were able to predict gene abundance levels of targeted genes with root mean squared error (RMSE) under one standard deviation of the ground truth levels. This result suggests that plants can be used as “probes” into the soil microbiome, and the hyperspectral sensing allows for a much higher throughput than traditional approaches.

However, the cost of the high-end hyperspectral sensor used to collect data in this study, detailed in <ramamoorthy2022> can be prohibitive for future users. It may be of great practical value if similar predictions can be made from less expensive data, such as lower-cost hyperspectral sensors that do not cover as large a region of the electromagnetic spectrum, or broadband sensors such as multispectral or visible-light (RGB) cameras. Similarly, if a small subset of the wavebands could give predictions with an acceptable error level, future researchers might be able to save effort by focusing on the wavebands that are most informative, or by using sensors to target such wavebands <moghimi2018>. These problems might be addressed by applying feature selection techniques to find such a subset of wavebands.

Hyperspectral data tends to be very wide, with many more columns than rows. As a result, it is common to use dimensionality reduction techniques during preprocessing <vaddi2024>. Additionally, many studies have previously been conducted on hyperspectral feature selection and extraction for agricultural tasks, but generalizability among studies is poor <hennessy2020>, necessitating recalculation of feature importance on a case-by-case basis.

There are many existing feature selection techniques, but all perform the same fundamental task: selecting a subset of the most important variables from a dataset. In contrast to feature extraction techniques (e.g., principal component analysis), in which a transformation is applied to the dataset, feature selection methods preserve the original variable space, which can provide better interpretability. Although subcategories of feature selection methods are not well-defined, most can be classified as filter methods, embedded methods, or wrapper methods.

Filter methods typically use a more traditional statistical approach to feature importance. In contrast with embedded and wrapper methods, these do not require models to be trained; these rely on properties of the dataset itself. Filter methods are typically applied during the preprocessing phase in a machine learning pipeline.

While filter methods do not require a model to be trained, embedded methods use inherent properties of certain model types to rank feature importance. For example, linear models assign a weight to each variable, which can be used to rank the variables’ relative importance. Similarly, random forest models have a variety of metrics (e.g., Gini impurity) corresponding to the importance of each node in the tree. While these methods require training models, and only specific models can use embedded methods, the models only need to be trained once.

Wrapper methods train a model on a subset of the data, evaluate the model’s performance, and modify the set of variables chosen. This process is repeated until convergence. Wrapper methods are model agnostic, which provide more flexibility than embedded methods, but the repeated retraining can be computationally expensive.

Because different feature selection and extraction methods can produce different results on the same data, it is common to compile features chosen by an ensemble of methods <hennessy2020>. For example, <fei2022> used an ensemble of three embedded methods – two based on random forests, and one based on LASSO – and one filter method to select the most relevant wavebands for predicting corn yield. Similarly, <elsherbiny2021> selected wavebands to predict rice canopy water content using an ensemble of embedded methods, extraction methods, and feature engineering. Ensemble approaches to hyperspectral feature selection can also be used for classification tasks, such as disease detection in peanuts <wei2021>.

**Methods**

**Subsection: Data**

The dataset used in this study came from the experiment described in <cite ramamoorthy2022> and extended in <brooks2024>. 80 cotton plants (*Gossypium hirsutum*) were grown in a greenhouse under controlled conditions. 5 observations of leaf hyperspectral reflectance per plant were measured with a PSR+3500 spectroradiometer (Spectral Evolution, Haverhill, MA, USA) on October 20, 2021. This had a spectral range of [350, 2500] nm with a spectral resolution of 2.8 nm at 700 nm, 8 nm at 1500 nm, and 6 nm at 2100 nm, full width at half maximum. This was resampled to give a consistent spectral resolution of 1 nm across the spectrum, resulting in 2151 wavebands. However, there were large regions of very high correlation across the spectrum (Figure <corr figure>), particularly in the infrared regions, making feature selection even more important in this domain.

In addition to the hyperspectral reflectance data, gene abundance levels were extracted from soil attached to the cotton plant roots, as described in <brooks2024>. Five genes were targeted: 16S rRNA <nadkarni2002>, 18S rRNA <liu2015>, ureC <koper2004>, phoA <han2012>, and cbbLR <selesi2005>. 16S rRNA and 18S rRNA have been shown to be indicators of soil health <cite …?>, and the other genes were chosen based on their roles in the carbon cycle. Lastly, the gene abundance levels were log10 transformed before use.

**Subsection: Overall workflow**

The workflow illustrated in Figure <workflow> was applied to each of the five genes. Two regions of the dataset were considered: one with all wavebands (350-2500 nm) as predictors, and one with visible light only (400-700 nm). The dataset was then split 80%/20% into training and testing sets, respectively. Each predictor and target variable was standardized according to the training set’s distribution to have a mean of 0 and standard deviation of 1, allowing for meaningful comparison among wavebands and gene levels.

We chose six waveband (feature) selection methods to consider, incorporating several paradigms. Mutual information and hierarchical clustering were both filter methods, using the linear model coefficients was an embedded method, and permutation importance and the genetic algorithms fit the wrapper paradigm. These five methods chose important wavebands from the entire region of consideration (either all wavebands or visible light only). Then, a second genetic algorithm selected a consensus of the top wavebands from the concatenated results of the previous five methods.

After each of the six methods produced a subset of wavebands deemed most important for predictions, we trained a new elastic net model on the training set, but only using that subset of features. This was intended as a validation test; for example, if a method’s feature choices were poor, the validation model’s error would be very high. In addition to the validation models, a baseline model using no feature selection methods was trained for each gene and region as a basis for comparison. The elastic net algorithm was chosen based on the results of <Brooks2024>, in which its predictions outperformed the random forest and LASSO algorithms on this dataset. Throughout the workflow, 5-fold grid search cross-validation was used within the training set for tuning the regularization penalty and ℓ1/ ℓ2 ratio hyperparameters. Each model was evaluated on the testing set based on the root mean square error (RMSE), mean absolute error (MAE), and R2 metrics.

Because the target variables were standardized, the performance of models predicting different genes could be compared on the same scale. Additionally, a “blind” model always predicting the mean value would be guaranteed a RMSE and MAE of approximately[[1]](#footnote-1) 1, the standard deviation. This provided a quantitative basis of comparison to evaluate model bias.

We implemented this workflow in Python 3.12.7 using the *scikit-learn* 1.5.2 <cite scikit-learn> and *sklearn-genetic-opt* 0.11.1 <cite genetic> libraries.

**Subsection: Filter methods**

One feature selection method used in this study used mutual information as a metric. Mutual information, equivalent to information gain in this context, quantifies the dependence of one variable given another. The mutual information of each waveband with the gene level was calculated, and the function returned the 64 wavebands with the largest MI values.

Our clustering method for waveband selection first calculated the Pearson correlation matrix for the predictors (wavebands). We then applied agglomerative clustering to the correlation matrix, using a distance threshold of 0.999 and Ward linkage. This threshold was chosen based on exploratory data analysis conducted in <brooks2024>, which demonstrated the high correlation of adjacent wavebands across the spectrum. After the clusters of wavebands were generated, one waveband was randomly chosen from each cluster. The set of these representative wavebands was returned as the clustering method’s selections. Because the clustering method only considered the similarity among wavebands, not taking the target variables into account, it only needed to be executed twice: once for the full dataset, and once for the visible-light-only version of the dataset.

**Subsection: Embedded method**

The fundamental task of linear regression models is to find the set of weights *w* that minimizes the error (or variance) of the equation over all data points. Several variants exist to minimize the unwanted effect of outliers in the dataset; one robust variant is elastic net regression, which finds the optimal values for *w* in the following expression:

where α is the regularization penalty, and ρ is the ratio between the ℓ1 and ℓ2 norms.

In <brooks2024>, one reason that elastic net was chosen for investigation was its inherent support of an embedded feature selection method, namely, using its coefficients. A variable with a large positive or negative coefficient will have a stronger effect on predictions than one with a coefficient close to zero. When using this method, our function returned the set of 64 wavebands with the highest absolute valued coefficients.

**Subsection: Wrapper methods**

Permutation importance is another method of determining the magnitude of effect a variable has on a model’s predictions. After training and testing a model on the data, a single variable is permuted over the number of observations. A new model is then trained on the modified dataset. If the subsequent performance metrics are worse than the first model’s, this indicates that the permuted variable was important to the model’s predictions. A greater decrease in prediction quality indicates a higher importance for the permuted variable. This process is repeated for all variables in the dataset. To run permutation importance variable selection, we first applied agglomerative clustering <Add cross reference to clustering> among the variables, which was intended to reduce overall correlation among features. After calculating permutation importance for all variables, the function selected the 64 wavebands with the highest scores.

The final method tested in this study used a genetic algorithm, implemented in the *sklearn-genetic-opt* Python library <cite genetic>, to select wavebands. A subset of variables was selected and iteratively improved according to an evolutionary algorithm. The effectiveness of each variable set was evaluated using 5-fold cross-validation on an elastic net model. The selection process minimized both the cross-validation RMSE and the number of features selected, with a maximum of 64 wavebands. When the genetic algorithm was used an additional time to find a consensus of the other waveband selections (see Figure <workflow>), a maximum of 16 wavebands were selected.

**Results and Discussion**

**Subsection: Validation model metrics**

As a caveat when interpreting validation model metrics: assuming a model has not overfit, reducing the number of variables will reduce the quality of prediction. In this study, different waveband selection methods produced different numbers of features; models with fewer features tended to have higher RMSE, and vice versa. As the number of features was manually set as a hyperparameter for some methods, this could be adjusted. The key results are the wavebands deemed important for each gene, provided that method’s error was reasonably similar to the baseline, and the validation model metrics should not be taken at face value.

As <cross reference method\_gene\_rmse> suggests, a validation model’s RMSE has more to do with the gene being predicted than the waveband selection method. For the same gene, different methods tended to have similar metrics. No waveband selection method resulted in a definitively superior or inferior RMSE for all five genes in this study. 16S rRNA and 18S rRNA are the more relevant genes in this dataset for monitoring soil health.

Elastic net models predicting 16S rRNA conclusively had the best results; the baseline model (using no waveband selection methods) using visible light wavebands only had the best metrics out of all models tested (RMSE = 0.724, MAE = 0.576, R2 = 0.501), outperforming the baseline model using all wavebands (RMSE = 0.762, MAE = 0.624, R2 = 0.448). This result not only demonstrates the feasibility of predicting 16S rRNA using hyperspectral data, but the better performance of the visible-light-only model also suggests the possibility of using more accessible sensors – hyperspectral sensors with a narrower range, multispectral sensors, or visible light sensors – in the future.

18S rRNA, on the other hand, was the most difficult gene to predict, with all validation models having RMSE exceeding 1, the standard deviation in the standardized set of gene levels. However, all models’ MAE scores were less than 1, with the best model selecting wavebands based on mutual information from the entire dataset (RMSE = 1.070, MAE = 0.860, R2 = 0.056). Although the difference between RMSE and MAE seems fairly large, it was typical for all genes’ validation models, suggesting 18S rRNA does not have an unusually large number of outliers compared to the other genes, and that the poor prediction quality could be due to the lack of a signal in the data.

Validation models predicting cbbLR, phoA, and ureC generally gave RMSE only slightly less than 1. This result is not as conclusive as the results of the other two genes, as this small improvement in prediction accuracy over a “blind” model does not suggest the presence of either a strong or weak signal in the hyperspectral data.

**Subsection: Waveband selections**

Even though the hyperspectral data had a spectral resolution of 1 nm, waveband selection at this level of precision has questionable generalizability. Wavebands tended to have very large Pearson correlations with neighboring wavebands for very large regions of the spectrum, particularly in the near infrared (NIR) region <cross reference to corr figure>. With correlation greater than 0.999 for adjacent wavebands across the spectrum, such wavebands become practically interchangeable. At this level, any variation in waveband selections might be due more to noise or stochastic algorithm choices than a true signal in the data. For this reason, we report waveband selection results at a coarser resolution, rounding to the nearest 10 nm.

Generally, there was not strong agreement for important wavebands among regions and genes, but a few patterns emerged. Broadly speaking, there was a preference for visible light when building models on all wavebands, and there was a preference for green, yellow, and orange light when building models on visible light wavebands only. Figures <consensus\_histogram> and <consensus\_histogram\_gene> show the relatively wide dispersal of wavebands selected by consensus methods over each region.

When allowing models to select from all wavebands, small clusters at 510-560 (green), 700-720 (RE), and 970-980 (NIR) were each chosen by multiple genes’ consensus models. The cluster from 1890-1940 (FSWIR) was almost entirely chosen by the 16S rRNA consensus model, whose validation predictions had among the best metrics out of all tested models.

The visible-only consensus models tended to favor orange and red light; the largest cluster had twin peaks at 620 (orange) and 630 (red). Several consensus methods also converged on 670 (red) and 700 (RE). To a lesser extent, green and yellow light wavebands were selected: 510 (green), 540 (green), and 580 (yellow) formed a second tier of peaks. While some genes’ consensus models selected wavebands from the region of purple light, this was less common.

In particular, Figure <bact\_baseline\_vis\_coeffs\_bw> shows the coefficients of the 16S rRNA visible-only baseline model, which had the best metrics overall. Consistent with the consensus model selections, features around 540 (green) and 590-600 (yellow-orange) were assigned high coefficients by the baseline model. In contrast with the consensus models for all genes, though, the 16S rRNA visible-only baseline model placed high importance on wavebands near 440 (purple). This region of the spectrum is not often found to be important in remote sensing applications with plants <hennessy2020>, so this bears further investigation. Similarly, the coefficients of the 16S rRNA visible-only consensus model <bact\_consensus\_vis\_coeffs\_bw> indicate a very similar contour to the coefficients in <bact\_baseline\_vis\_coeffs\_bw>, even though the consensus model had two layers of feature selection applied. This agreement strengthens the results of both, especially considering both models’ superior metrics.

**Subsection: Limitations and future work**

The largest limitation of this study is the small size of the dataset; after all, costliness of data collection in this domain is the central motivation of this study. While the methods employed in this study were chosen to accommodate the small dataset, replicating the analysis on a larger dataset can strengthen confidence in its results by making the models more robust. Additionally, further work needs to be done to generalize this study’s methodology to predict other genes, to measure plants other than cotton, and most importantly, to predict gene abundance levels in actual field conditions.

In this study, we manually reduced the spectral range to visible light to simulate data collected with more accessible sensors. Surprisingly, this resulted in the model with the best metrics (16S rRNA, visible-only, baseline). Future studies might similarly investigate the effect of coarser spectral resolutions on model prediction quality.

The waveband selection workflow in this study was designed to reduce the number of wavebands selected at each step, but the number of features selected was ultimately arbitrary. We reasoned that, since in practice it would be trivial to use the entire dataset as input to predict gene abundance levels, the objective was not necessarily to select features to create a model with the best metrics, but to investigate which wavebands had the strongest effect when making these predictions. If future studies shift emphasis to optimize prediction quality, it may be useful to add the number of wavebands selected as an additional hyperparameter and adjust the tuning algorithms accordingly.

**Conclusion**

Monitoring the soil microbiome through laboratory tests is laborious and costly. Instead, predicting soil gene abundance levels using hyperspectral reflectance of leaves of cotton planted in that soil shows promise in increasing throughput. Our elastic net models showed great promise in predicting 16S rRNA. ~~An elastic net model using all the wavebands was able to predict the standardized values of 16S rRNA with RMSE = 0.762, MAE = 0.624, and R~~~~2~~ ~~= 0.448.~~ Restricting the wavebands to visible light only resulted in better metrics than when the model included the infrared reflectance, suggesting that more accessible sensors could be used in future studies for this task. Different methods predicted ureC, phoA, and cbbLR slightly below the standard deviation, but 18S rRNA was not predictable using this methodology.

We used an ensemble of feature selection techniques to determine regions of the electromagnetic spectrum important for predicting five gene levels. No feature selection method consistently performed better or worse than the others for all genes. Wavebands in the visible region – particularly green, yellow, and orange light – were most favored by the feature selection methods for all five genes. However, the superior 16S rRNA models also assigned higher importance to the purple and FSWIR regions. This study demonstrates the feasibility of predicting soil 16S rRNA abundance values using only cotton hyperspectral reflectance, and the results suggest the possibility of using more accessible sensors such as multispectral or RGB cameras in the future.

**Data availability**

The data, source code, results, and analysis for this study can be accessed at https://github.com/rockerd1134/SoilGeneHS.

**Acknowledgments**

This research used resources provided by the SCINet project and/or the AI Center of Excellence of the USDA Agricultural Research Service, ARS project numbers 0201-88888-003-000D and 0201-88888-002-000D.

This work did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

**References**

<TODO>

1. Although these values should be exactly equal to 1 assuming the training and testing sets have identical distributions, this is not the case in practice. [↑](#footnote-ref-1)