**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
    - Disclaimer: Even though the hyperspectral data had a spectral resolution of 1 nm, waveband selection <TODO: finish this>
    - Generally no strong agreement for important wavebands among methods and genes, but a few weak patterns
    - 16S
* 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**

<Intro here. Should be background + lit review>

**Methods**

**Subsection: Data**

* Data
  + 400 cotton plants, greenhouse controlled environment <Should experimental design be discussed here, since the different treatments were only relevant for this study to the extent they generated variation in the data?>
  + Hyperspectral reflectance measurements, range of [350,2500] nm, spectral resolution of 1 nm (double check that this is accurate, or if it were preprocessed by GRI or something before it got to me)
  + qPCR data <I’ll write this based on prev manu, but it definitely needs them to check>

**Subsection: Overall workflow**

The workflow illustrated in Figure <workflow> was applied to each of the five genes. Two versions 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. 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 validation model was evaluated on the testing set based on the root mean square error (RMSE), mean absolute error (MAE), and R2 metrics. <TODO: add sentence here about how the targets were standardized, so SD = 1 = RMSE = MAE of blind model> Because the target variables were also standardized,

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 main reason that elastic net models were chosen in <brooks2024> was their inherent support for embedded feature selection. Because elastic net models are just variants of linear regression models, each input variable is assigned a weight. The absolute values of

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**

<Incorporates both sections>

**Conclusion**

<Conclusion here>

**Data availability**

The data and source code for this study can be accessed at <https://github.com/rockerd1124/DirtSpectra>. <Update this when the repo name changes>

**Acknowledgements**

<This section optional. Put SCINet here?> This research used resources provided by the SCINet projet 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**

<From EndNote. If other authors have any necessary references, probably easiest to send to me so I can put in my EndNote and export all together>