Paper Title\* (use style: *paper title*)

\*Note: Sub-titles are not captured in Xplore and should not be used

line 1: 1st Given Name Surname   
line 2: *dept. name of organization (of Affiliation)*  
line 3: *name of organization (of Affiliation)*line 4: City, Country  
line 5: email address or ORCID  
  
line 1: 2nd Given Name Surname  
line 2: *dept. name of organization (of Affiliation)*  
line 3: *name of organization (of Affiliation)*line 4: City, Country  
line 5: email address or ORCID  
  
line 1: 3rd Given Name Surname  
line 2: *dept. name of organization (of Affiliation)*  
line 3: *name of organization (of Affiliation)*line 4: City, Country  
line 5: email address or ORCID  
  
line 1: 4th Given Name Surname  
line 2: *dept. name of organization (of Affiliation)*  
line 3: *name of organization (of Affiliation)*line 4: City, Country  
line 5: email address or ORCID

line 1: 5th Given Name Surname  
line 2: *dept. name of organization (of Affiliation)*  
line 3: *name of organization (of Affiliation)*line 4: City, Country  
line 5: email address or ORCID

line 1: 6th Given Name Surname  
line 2: *dept. name of organization (of Affiliation)*  
line 3: *name of organization (of Affiliation)*line 4: City, Country  
line 5: email address or ORCID

*Abstract*—This electronic document is a “live” template and already defines the components of your paper [title, text, heads, etc.] in its style sheet. *\*CRITICAL: Do Not Use Symbols, Special Characters, Footnotes, or Math in Paper Title or Abstract*. (*Abstract*)

Keywords—component, formatting, style, styling, insert (key words)

# Introduction (*Heading 1*)

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.

<TODO: add sentence here about how the targets were standardized, so SD = 1 = RMSE = MAE of blind model> 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, 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 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 y=mX+b 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:

\frac{1}{2n\_{samples}}\left|\left|Xw-y\right|\right|\_2^2+\alpha\rho\left|\left|w\right|\right|\_1+\frac{\alpha\left(1-\rho\right)}{2}\left|\left|w\right|\right|\_2^2

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 <TODO: cite here, probably Hennessy but check others>, 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..Ease of Use

## Selecting a Template (Heading 2)

First, confirm that you have the correct template for your paper size. This template has been tailored for output on the US-letter paper size. If you are using A4-sized paper, please close this file and download the file “MSW\_A4\_format”.

## Maintaining the Integrity of the Specifications

Identify applicable funding agency here. If none, delete this text box.

The template is used to format your paper and style the text. All margins, column widths, line spaces, and text fonts are prescribed; please do not alter them. You may note peculiarities. For example, the head margin in this template measures proportionately more than is customary. This measurement and others are deliberate, using specifications that anticipate your paper as one part of the entire proceedings, and not as an independent document. Please do not revise any of the current designations.

# Prepare Your Paper Before Styling

Before you begin to format your paper, first write and save the content as a separate text file. Complete all content and organizational editing before formatting. Please note sections A-D below for more information on proofreading, spelling and grammar.

Keep your text and graphic files separate until after the text has been formatted and styled. Do not use hard tabs, and limit use of hard returns to only one return at the end of a paragraph. Do not add any kind of pagination anywhere in the paper. Do not number text heads-the template will do that for you.

## Abbreviations and Acronyms

Define abbreviations and acronyms the first time they are used in the text, even after they have been defined in the abstract. Abbreviations such as IEEE, SI, MKS, CGS, sc, dc, and rms do not have to be defined. Do not use abbreviations in the title or heads unless they are unavoidable.

## Units

* Use either SI (MKS) or CGS as primary units. (SI units are encouraged.) English units may be used as secondary units (in parentheses). An exception would be the use of English units as identifiers in trade, such as “3.5-inch disk drive”.
* Avoid combining SI and CGS units, such as current in amperes and magnetic field in oersteds. This often leads to confusion because equations do not balance dimensionally. If you must use mixed units, clearly state the units for each quantity that you use in an equation.
* Do not mix complete spellings and abbreviations of units: “Wb/m2” or “webers per square meter”, not “webers/m2”. Spell out units when they appear in text: “. . . a few henries”, not “. . . a few H”.
* Use a zero before decimal points: “0.25”, not “.25”. Use “cm3”, not “cc”. (*bullet list*)

## Equations

The equations are an exception to the prescribed specifications of this template. You will need to determine whether or not your equation should be typed using either the Times New Roman or the Symbol font (please no other font). To create multileveled equations, it may be necessary to treat the equation as a graphic and insert it into the text after your paper is styled.

Number equations consecutively. Equation numbers, within parentheses, are to position flush right, as in (1), using a right tab stop. To make your equations more compact, you may use the solidus ( / ), the exp function, or appropriate exponents. Italicize Roman symbols for quantities and variables, but not Greek symbols. Use a long dash rather than a hyphen for a minus sign. Punctuate equations with commas or periods when they are part of a sentence, as in:

*a**b* 

Note that the equation is centered using a center tab stop. Be sure that the symbols in your equation have been defined before or immediately following the equation. Use “(1)”, not “Eq. (1)” or “equation (1)”, except at the beginning of a sentence: “Equation (1) is . . .”

## Some Common Mistakes

* The word “data” is plural, not singular.
* The subscript for the permeability of vacuum **0, and other common scientific constants, is zero with subscript formatting, not a lowercase letter “o”.
* In American English, commas, semicolons, periods, question and exclamation marks are located within quotation marks only when a complete thought or name is cited, such as a title or full quotation. When quotation marks are used, instead of a bold or italic typeface, to highlight a word or phrase, punctuation should appear outside of the quotation marks. A parenthetical phrase or statement at the end of a sentence is punctuated outside of the closing parenthesis (like this). (A parenthetical sentence is punctuated within the parentheses.)
* A graph within a graph is an “inset”, not an “insert”. The word alternatively is preferred to the word “alternately” (unless you really mean something that alternates).
* Do not use the word “essentially” to mean “approximately” or “effectively”.
* In your paper title, if the words “that uses” can accurately replace the word “using”, capitalize the “u”; if not, keep using lower-cased.
* Be aware of the different meanings of the homophones “affect” and “effect”, “complement” and “compliment”, “discreet” and “discrete”, “principal” and “principle”.
* Do not confuse “imply” and “infer”.
* The prefix “non” is not a word; it should be joined to the word it modifies, usually without a hyphen.
* There is no period after the “et” in the Latin abbreviation “et al.”.
* The abbreviation “i.e.” means “that is”, and the abbreviation “e.g.” means “for example”.

An excellent style manual for science writers is [7].

# Using the Template

After the text edit has been completed, the paper is ready for the template. Duplicate the template file by using the Save As command, and use the naming convention prescribed by your conference for the name of your paper. In this newly created file, highlight all of the contents and import your prepared text file. You are now ready to style your paper; use the scroll down window on the left of the MS Word Formatting toolbar.

## Authors and Affiliations

**The template is designed for, but not limited to, six authors.** A minimum of one author is required for all conference articles. Author names should be listed starting from left to right and then moving down to the next line. This is the author sequence that will be used in future citations and by indexing services. Names should not be listed in columns nor group by affiliation. Please keep your affiliations as succinct as possible (for example, do not differentiate among departments of the same organization).

### For papers with more than six authors: Add author names horizontally, moving to a third row if needed for more than 8 authors.

### For papers with less than six authors: To change the default, adjust the template as follows.

#### Selection: Highlight all author and affiliation lines.

#### Change number of columns: Select the Columns icon from the MS Word Standard toolbar and then select the correct number of columns from the selection palette.

#### Deletion: Delete the author and affiliation lines for the extra authors.

## Identify the Headings

Headings, or heads, are organizational devices that guide the reader through your paper. There are two types: component heads and text heads.

Component heads identify the different components of your paper and are not topically subordinate to each other. Examples include Acknowledgments and References and, for these, the correct style to use is “Heading 5”. Use “figure caption” for your Figure captions, and “table head” for your table title. Run-in heads, such as “Abstract”, will require you to apply a style (in this case, italic) in addition to the style provided by the drop down menu to differentiate the head from the text.

Text heads organize the topics on a relational, hierarchical basis. For example, the paper title is the primary text head because all subsequent material relates and elaborates on this one topic. If there are two or more sub-topics, the next level head (uppercase Roman numerals) should be used and, conversely, if there are not at least two sub-topics, then no subheads should be introduced. Styles named “Heading 1”, “Heading 2”, “Heading 3”, and “Heading 4” are prescribed.

## Figures and Tables

#### Positioning Figures and Tables: Place figures and tables at the top and bottom of columns. Avoid placing them in the middle of columns. Large figures and tables may span across both columns. Figure captions should be below the figures; table heads should appear above the tables. Insert figures and tables after they are cited in the text. Use the abbreviation “Fig. 1”, even at the beginning of a sentence.

1. Table Type Styles

| Table Head | Table Column Head | | |
| --- | --- | --- | --- |
| Table column subhead | Subhead | Subhead |
| copy | More table copya |  |  |

1. Sample of a Table footnote. (*Table footnote*)
2. Example of a figure caption. (*figure caption*)

Figure Labels: Use 8 point Times New Roman for Figure labels. Use words rather than symbols or abbreviations when writing Figure axis labels to avoid confusing the reader. As an example, write the quantity “Magnetization”, or “Magnetization, M”, not just “M”. If including units in the label, present them within parentheses. Do not label axes only with units. In the example, write “Magnetization (A/m)” or “Magnetization {A[m(1)]}”, not just “A/m”. Do not label axes with a ratio of quantities and units. For example, write “Temperature (K)”, not “Temperature/K”.

##### Acknowledgment *(Heading 5)*

The preferred spelling of the word “acknowledgment” in America is without an “e” after the “g”. Avoid the stilted expression “one of us (R. B. G.) thanks ...”. Instead, try “R. B. G. thanks...”. Put sponsor acknowledgments in the unnumbered footnote on the first page.

##### References

The template will number citations consecutively within brackets [1]. The sentence punctuation follows the bracket [2]. Refer simply to the reference number, as in [3]—do not use “Ref. [3]” or “reference [3]” except at the beginning of a sentence: “Reference [3] was the first ...”

Number footnotes separately in superscripts. Place the actual footnote at the bottom of the column in which it was cited. Do not put footnotes in the abstract or reference list. Use letters for table footnotes.

Unless there are six authors or more give all authors’ names; do not use “et al.”. Papers that have not been published, even if they have been submitted for publication, should be cited as “unpublished” [4]. Papers that have been accepted for publication should be cited as “in press” [5]. Capitalize only the first word in a paper title, except for proper nouns and element symbols.

For papers published in translation journals, please give the English citation first, followed by the original foreign-language citation [6].

1. G. Eason, B. Noble, and I. N. Sneddon, “On certain integrals of Lipschitz-Hankel type involving products of Bessel functions,” Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955. *(references)*
2. J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.
3. I. S. Jacobs and C. P. Bean, “Fine particles, thin films and exchange anisotropy,” in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.
4. K. Elissa, “Title of paper if known,” unpublished.
5. R. Nicole, “Title of paper with only first word capitalized,” J. Name Stand. Abbrev., in press.
6. Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, “Electron spectroscopy studies on magneto-optical media and plastic substrate interface,” IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].
7. M. Young, The Technical Writer’s Handbook. Mill Valley, CA: University Science, 1989.
8. K. Eves and J. Valasek, “Adaptive control for singularly perturbed systems examples,” Code Ocean, Aug. 2023. [Online]. Available: <https://codeocean.com/capsule/4989235/tree>
9. D. P. Kingma and M. Welling, “Auto-encoding variational Bayes,” 2013, arXiv:1312.6114. [Online]. Available: <https://arxiv.org/abs/1312.6114>
10. S. Liu, “Wi-Fi Energy Detection Testbed (12MTC),” 2023, gitHub repository. [Online]. Available: https://github.com/liustone99/Wi-Fi-Energy-Detection-Testbed-12MTC
11. “Treatment episode data set: discharges (TEDS-D): concatenated, 2006 to 2009.” U.S. Department of Health and Human Services, Substance Abuse and Mental Health Services Administration, Office of Applied Studies, August, 2013, DOI:10.3886/ICPSR30122.v2

**IEEE conference templates contain guidance text for composing and formatting conference papers. Please ensure that all template text is removed from your conference paper prior to submission to the conference. Failure to remove template text from your paper may result in your paper not being published.**