*Supplementary information for*

**Pollen Clustering Strategies Using a Newly Developed Single-Particle Fluorescence Spectrometer**

Benjamin E. Swanson and J. Alex Huffman

University of Denver, Department of Chemistry and Biochemistry, Denver CO 80210, USA

*Correspondence to: J. Alex Huffman (*[*alex.huffman@du.edu*](mailto:alex.huffman@du.edu)*)*

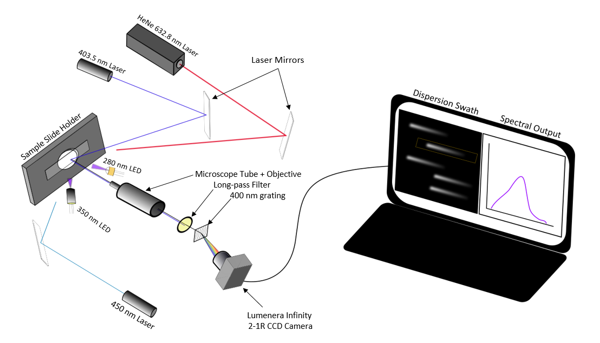
****

Figure S1: Schematic of instrumental design and operation.

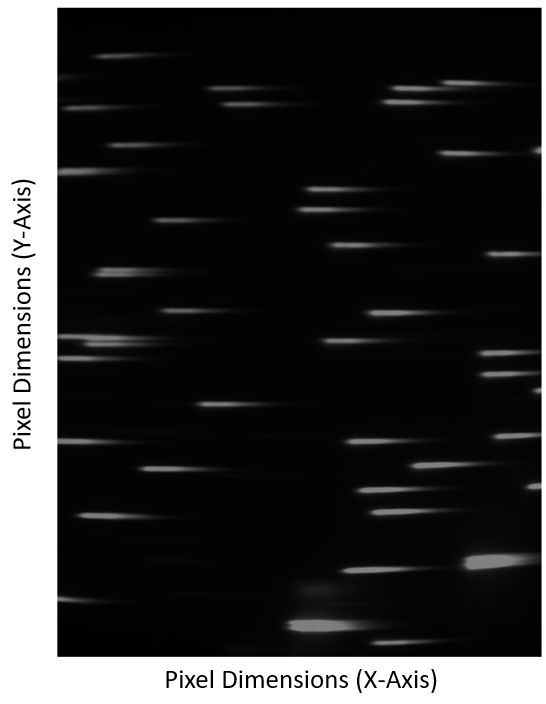


Figure S2: Camera viewing area with approximately 50 visible particle signals, each represented as a swath (spectrum) of dispersed fluorescent light from ~400 to 700 nm (left to right).

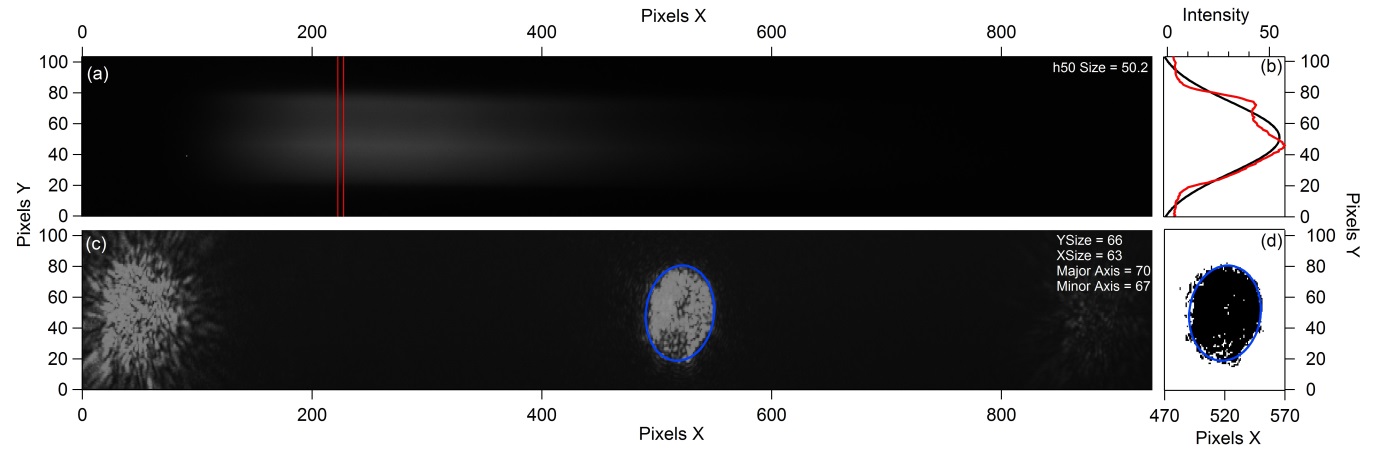


Figure S3: Particle sizing figure, associated with Figure 1 in the main text (where images in (a) and (c) were chopped for visual clarity).

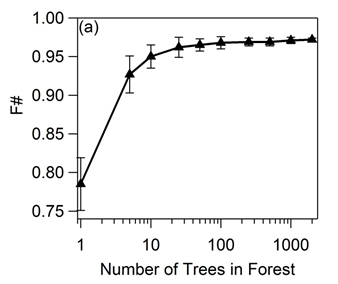


Figure S4: Analysis of the Random Forest model, showing increase in accuracy (*F*) as the number of trees are increased. Triangles represent averages of 5 trials, and vertical bars represent the standard deviation of those trials.

C:\Users\benjamin.swanson\Desktop\Loss Function.tiff

Figure S5: Exponential loss plot for GB model. Black trace indicates the reduction of error as subsequent trees are developed. Green trace represents reduction in error after *k* folds of the cross-validation test sets. Blue dotted line indicates minimum point on green trace.

The blue dotted line represents the last iteration in the model that does not overfit data. As the model moves past this iteration (tree 98), the data becomes more likely to overfit, preventing new observations from being accurately predicted. Past this point the test loss curve (green) begins to diverge upwards, away from the training curve (black), visually representing overfitting.

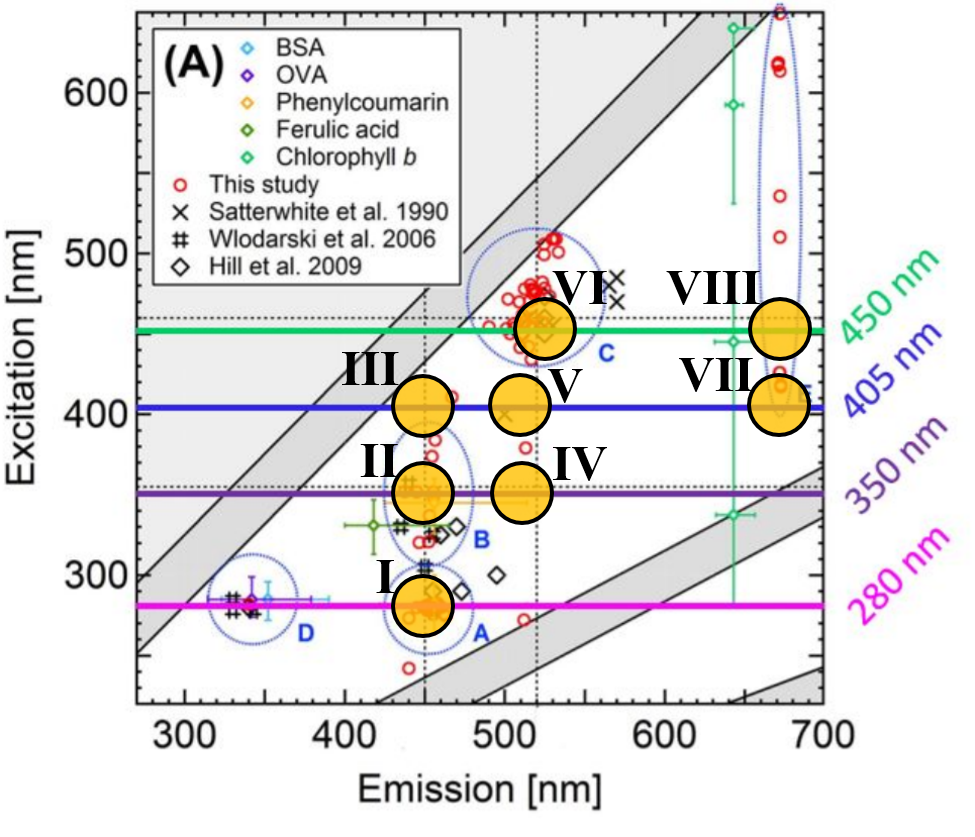
****

Figure S6: Summary of previous work on pollen fluorescence, shown as an excitation emission matrix, with the four excitation wavelengths used in the current version of our instrument represented as colored horizontal lines. Commonly observed fluorophores modes are highlighted as orange circles, with numeric assigned as defined in Section 3.1 of the main text. Figure adapted from Pöhlker et al., 2013.

Citation: Pöhlker, C., Huffman, J.A., and Pöschl, U. (2013). Autofluorescence of atmospheric bioaerosols: Spectral fingerprints and taxonomic trends of pollen. *Atmos. Meas. Tech.,* 6(12):3369–3392, <https://doi.org/10.5194/amt-6-3369-2013>.

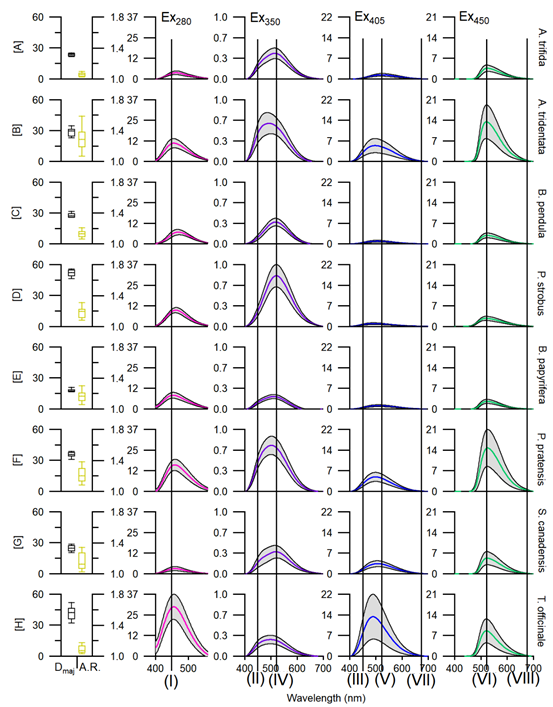


Figure S7: Particle size and spectral characteristics of the eight pollen species examined. Analogous to Figure 3 in main text, with black vertical lines added to represent fluorophores emission modes as defined in Section 3.1 (see also Fig. S6). Pollen species name listed on right of each panel.

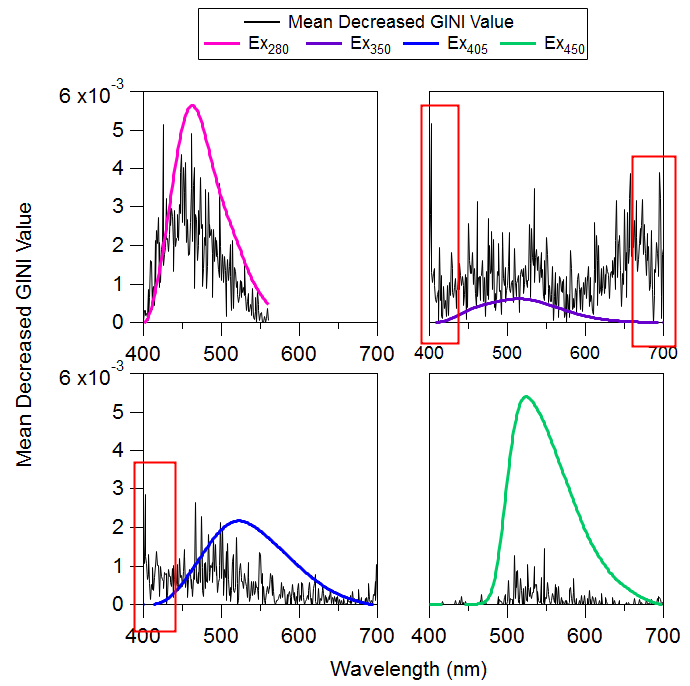


Figure S8: Model importance values (black) plotted with the reference spectra (color). Areas of the spectrum with high importance (Mean Decreased Gini values) associated with spectral noise (image background light) are boxed in red. Emission spectra at wavelengths associated with these areas of high noise were removed before input into classification so as to reduce model development based on noise. Figure S8 is analogous to Figure 7 in the main text, where portions of the black curves within the red boxes have already been removed (“correction” shown in Figure S9).

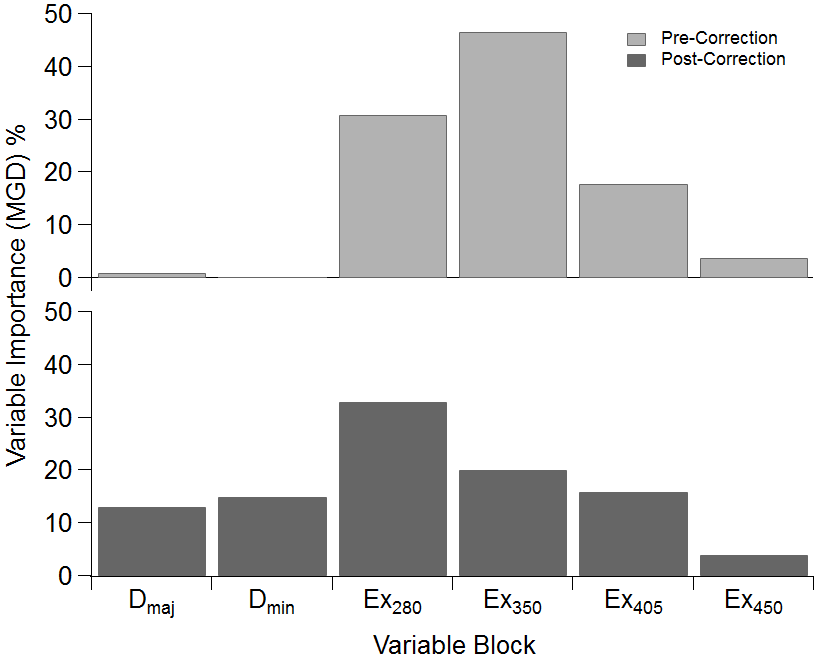


Figure S9: Changes in variable importance after removing spectral noise (see Fig. S8) and size-weighting in the gradient boosting system.

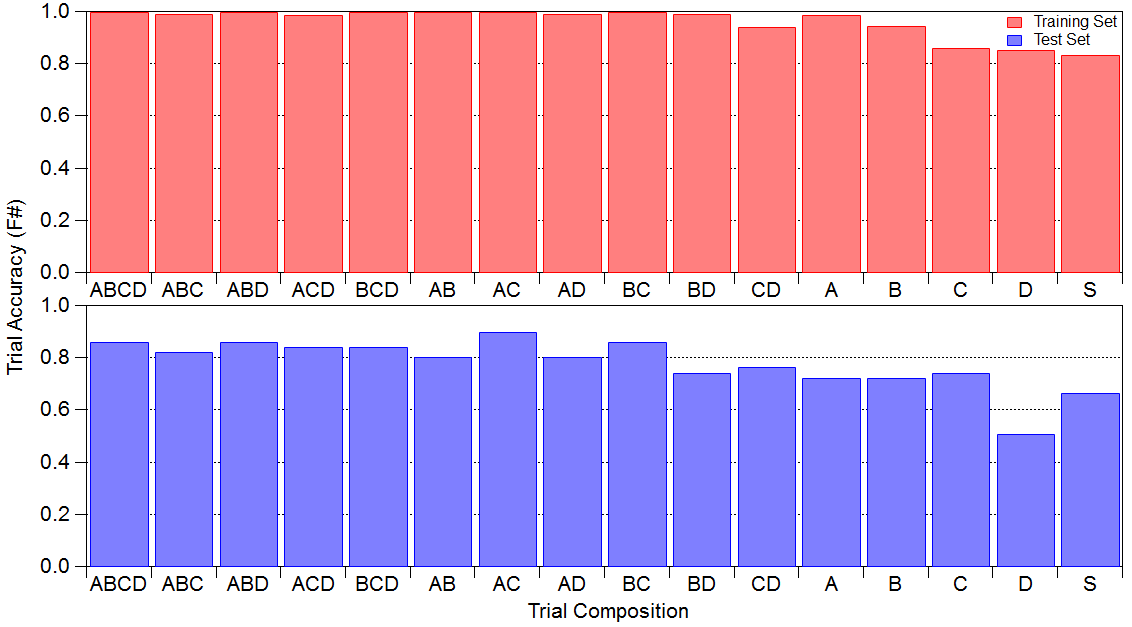
****

Figure S10: Accuracy of the RF algorithm following fifteen combinations of input variables. Excitation sources represented here as (A) 280 nm, (B) 350 nm, (C), 405 nm, and (D) 450 nm. All trials consist of a subset of 25% of the particle spectral data predicted to training models from 75% of the data. The final column represents input of no emission data, but only sizing (S) variables. Sizing data was not input into classification models except for tests associated with the last column.