1-1-2019

Development and Characterization of an Inexpensive Single-Particle Fluorescence Spectrometer for Detection and Classification of Pollen and Other Bioaerosols

Benjamin E. Swanson

University of Denver

Development and Characterization of an Inexpensive Single-Particle Fluorescence Spectrometer for Detection and Classification of Pollen and Other Bioaerosols

Abstract

Atmospheric aerosols are ubiquitous throughout the Earth's atmosphere and can be important with respect to environmental systems and human health. Pollen particles are a class of primary biological aerosol particles (PBAPs) that cost the United States billions of dollars a year in loss of productivity and healthcare costs due to allergy and respiratory effects. Traditional methods of pollen detection rely on collection and subsequent identification by visual microscopy, yet few measurement stations exist in the United States. As such, current pollen forecasting models have relatively high prediction uncertainty, especially in regions without sampling stations. Recently, laser-induced fluorescence instrumentation has been applied as one method to bridge gaps in bioaerosol detection and classification, though this instrumentation suffers from prohibitively high cost or analysis barriers.

This thesis describes the development, characterization, and preliminary application of a new single-particle fluorescence spectrometer geared towards bioaerosol, particularly pollen, analysis. A sequence of four laser or LED sources are used to excite the particles, which emit fluorescent light that is magnified then diffracted through a transmission grating into a simple digital camera. This instrument operates similar to a traditional spectroscope, though is able to collect spectral light from several small particles simultaneously. This process allows for spectroscopic analysis of many particles at the same time. The instrument went through several phases of both development and characterization. Development included the addition of several new excitation sources (two light-emitting diodes and one laser) to expand the number of fluorophores probed. A monochrome camera was also added to the system to circumvent issues caused by inexpensive point-and-shoot cameras. Methods to size the particles, as well as calibrations for camera parameters and systemic defects were also implemented. For defects in the optical surface and differences in source intensity, a spatial interpolation map was developed that reduces the error of identical particles depending on their location on the CCD from 17% to 3%.

Utilizing these techniques, four clustering and classification methods were examined with 8 species of commercial pollen in Chapter 4. The random forest (RF) and gradient boosting algorithms performed exceptionally well, both classifying above 95% accuracy. The RF technique was examined further due to computational advantages. Testing on source reduction revealed that the 405 and 450 nm sources were less important in classification models, with the latter having particularly low (3%) importance.

The classification techniques were utilized on freshly collected pollen standards in Chapter 5. 34 types of pollen were collected and classified to 90% accuracy at the species level. Pollen was also classified by species, allergenicity, as well as by plant type depending on their collection months, with one scenario being classified at 98% accuracy. A proof-of-concept was also provided for the prediction of new, ambient pollen samples to a developed random forest classification model from standard collections, in which several particles collected in a central location of the Botanic Gardens were classified as a type of tree that was seen to be pollinating on the same day.

Document Type

Dissertation

Degree Name

Ph.D.

Department
Chemistry and Biochemistry

First Advisor
John A. Huffman, Ph.D.

Second Advisor
Mark Siemens, Ph.D.

Keywords
Atmospheric, Bioaerosols, Clustering, Instrumentation, Laser-induced fluorescence, Pollen

Subject Categories
Chemistry | Physical Sciences and Mathematics

Development and Characterization of an Inexpensive Single-Particle Fluorescence Spectrometer for Detection and Classification of Pollen and Other Bioaerosols

---

A Dissertation

Presented to

the Faculty of Natural Sciences and Mathematics

University of Denver

In Partial Fulfullment

of the Requirements for the Degree

Doctor of Philosophy

---

by

Benjamin E. Swanson

November 2019

Advisor: J. Alex Huffman

Author: Benjamin E. Swanson
Title: Development and Characterization of an Inexpensive Single-Particle Fluorescence Spectrometer for Detection and Classification of Pollen and Other Bioaerosols
Adviser: J. Alex Huffman
Degree Date: November 2019

Abstract

Atmospheric aerosols are ubiquitous throughout the Earth's atmosphere and can be important with respect to environmental systems and human health. Pollen particles are a class of primary biological aerosol particles (PBAPs) that cost the United States billions of dollars a year in loss of productivity and healthcare costs due to allergy and respiratory effects. Traditional methods of pollen detection rely on collection and subsequent identification by visual microscopy, yet few measurement stations exist in the United States. As such, current pollen forecasting models have relatively high prediction uncertainty, especially in regions without sampling stations. Recently, laser-induced fluorescence instrumentation has been applied as one method to bridge gaps in bioaerosol detection and classification, though this instrumentation suffers from prohibitively high cost or analysis barriers.

This thesis describes the development, characterization, and preliminary application of a new single-particle fluorescence spectrometer geared towards bioaerosol, particularly pollen, analysis. A sequence of four laser or LED sources are used to excite the particles, which emit fluorescent light that is magnified then diffracted through a transmission grating into a simple digital camera. This instrument operates similar to a traditional spectroscope, though is able to collect spectral light from several small particles simultaneously. This process allows for spectroscopic analysis of many particles at the same time. The instrument went through several phases of both development and characterization. Development included the addition of several new excitation sources (two light-emitting diodes and one laser) to expand the number of fluorophores probed. A monochrome camera was also added to the system to circumvent issues caused by inexpensive point-and-shoot cameras. Methods to size the particles, as well as calibrations for camera parameters and systemic defects were also implemented. For defects in the optical surface and differences in source intensity, a spatial interpolation map was developed that reduces the error of identical particles depending on their location on the CCD from 17% to 3%.

Utilizing these techniques, four clustering and classification methods were examined with 8 species of commercial pollen in Chapter 4. The random forest (RF) and gradient boosting algorithms performed exceptionally well, both classifying above 95% accuracy. The RF technique was examined further due to computational advantages. Testing on source reduction revealed that the 405 and 450 nm sources were less important in classification models, with the latter having particularly low (3%) importance.

The classification techniques were utilized on freshly collected pollen standards in Chapter 5. 34 types of pollen were collected and classified to 90% accuracy at the species level. Pollen was also classified by species, allergenicity, as well as by plant type depending on their collection months, with one scenario being classified at 98% accuracy. A proof-of-concept was also provided for the prediction of new, ambient pollen samples to a developed random forest classification model from standard collections, in which several particles collected in a central location of the Botanic Gardens were classified as a type of tree that was seen to be pollinating on the same day.

Acknowledgements

I want to extend my gratitude towards my advisor and mentor, Dr. Alex Huffman. His guideance, constructive criticism, and critically thinking mind has helped shape me into the scientist I am today. I am especially thankful that he has respected the lives of his students and is willing to put forth the importance of balancing life and work. I appreciate the amount of space and time he was willing to give to allow me to grow as a scientist. I would like to acknowledge members, current and former, of the Huffman group including Amani Alhalwani, Rachel Davey, Marie Gosselin, and Nicole Savage. A special thanks to Sam Scherer and Samir Rezgui for the work they performed under my guideance. I am also extremely grateful for my collaborator, Dr. Donald Huffman of the University of Arizona, who developed the original concept for the instrument described here, and who's input, and help, has been extremely insightful and valuable. I also am incredibly thankful for the Department of Chemistry and Biochemistry for providing me with so much financial support through the Phillipson Graduate Fellowship from 2015-2019.

Personally, I would like to thank my wonderfully patient and loving wife Jecca, who without I would not be in the position I am today. Throughout the past ten years, she has helped me grow, both mentally and physically, and consistently undersells her contribution to the person I am today. Similarly, my family, Dana, Stacey, Max, and Griffin Swanson have done the same, and I cannot possibly thank them enough. My parents have been some of the most supportive people I have ever known, and I could not imagine how different my life would be without them. My family has my deepest appreciation and respect, and their support has been, quite literally, immesurable.

Table of Contents

Chapter One: Introduction

Chapter Two: Design and Basic Instrumental Operation of a Newly Developed Single-Particle Fluorescence Spectrometer

Chapter Three: Benchtop Model – Improvements on Design and Analysis

Chapter Four: Clustering Strategies for Pollen Classification

Chapter Five: Pollen Classification with a Recently Developed Fluorescence Spectrometer

Chapter Six: Conclusions

List of Tables

Chapter 4

Chapter 5

Appendix B

List of Figures

Chapter 1

Chapter 2

Chapter 3

Chapter 4

Chapter 5

Appendix A

Appendix B

Appendix D

Chapter One: Introduction

1.1 Primary Biological Aerosol Overview

Aerosols are small airborne particles that are small enough to be suspended in the air and are ubiquitous throughout the Earth's atmosphere. Large aerosols, known as coarse mode particles (i.e. > 2.5 μm in diameter), include material such as dust elevated mechanically into the air, sea-salt spray from ocean waves, and pollen ejected from trees. A sub-class of these aerosols are biological particles. These are often called primary biological aerosols (PBAPs), ejected directed from biological sources, e.g. pollen and fungal spores, as well as particles like plant fragments that are mechanically elevated into the atmosphere by other sources (Fröhlich-Nowoisky et al. 2016; Després et al. 2012).

PBAPs have been observed to have a wide variety of influence on human life and climate. For example, bioaerosols can transport airborne disease, induce allergies, and it has been proposed that they can influence a number of environmental systems (Fröhlich-Nowoisky et al. 2016; Douwes et al. 2003). Exposure due to important classes or species of bioaerosols has become an important area of study at places like composting facilities (Wéry 2014; Hryhorczuk et al. 2001; Bünger et al. 2000) and livestock farms (Wéry 2014; Millner 2009; Mackiewicz 1998) due to the elevated risk of exposure to bacteria, fungal spores, and toxins that may be harmful to the body. Many pollen types are well-known allergens that up to 30% of U.S. adults suffer from seasonally (Sofiev and Bergmann 2013; Peden and Reed 2010).

Bioaerosols represent a wide class of particles that can have a myriad of effects on many different systems, and their physical properties are less explored compared to other atmospheric aerosols (Morris et al. 2011; Ariya et al. 2009). Physical properties of the particles are extremely important in the identification of bioaerosols. PBAPs cover an extremely wide array of sizes, ranging from tens of nanometers to several hundred microns (Bartlett 2008; Pöschl 2005; Górny et al. 1999). As a result, size and shape measurements of a particle can be used to classify them into broad groupings. Pollen, the primary focus of this dissertation work, are supermicron particles (e.g. 4-100 μm) though fragmentation of pollen, and other supermicron PBAPs, can occur in the atmosphere under certain conditions (Taylor et al. 2007; Green et al. 2006; Górny et al. 2002).

1.2 Pollen Overview

1.2.1 Definition and Brief Biology

Pollen are microscopic grains released from the male portion of pollen-producing plants. Pollen grains are typically spherical or elliptical in shape and range between sizes of 4 and 100 μm, depending on the species (Bennett and Willis 2002; Leuschner 1993). These sizes can translate into different ejection and transport features. Conifer pollen, for example, tends to be large and have features such as air bladders, which are air-filled structures that increase surface area to help increase the ability of the pollen to stay aloft (Schwendemann et al. 2007). This allows conifer pollen to be ejected directly from the tree, inducing pollination after being carried through the air to another tree through a process referred to as anemophily. Broadleaf pollen is produced in the stamen of the flower and is generally stickier. Most pollen grains are surrounded by protective material, a hard outer exine shell, to help reduce the impact of environmental conditions. Stickier pollen includes a layer made of lipid and carotenoid compounds, called pollenkitt, on the surface (Pacini and Hesse 2005; Runions and Owens 2002). Pollen with larger amounts of pollenkitt are generally pollinated by animals, referred to as zoophily, with a majority pollinated by insects, known as entomophily. Pollenkitt is not included in plant species that utilize plant or animal transport, however, as both types of pollen lie on a scale where pollenkitt inclusion correlates with entomophilous behavior (Hesse 1981). Entomophilous species tend to be characterized by highly-developed structures with protrusions (echinate) on the pollen grains, helping to provide a natural barrier as well as increase the ability to stick to pollinators (Tanaka et al. 2004). Anemophilous pollen tends to possess psilate, or smooth surfaced, grains or exhibit a reticulate structure. If echinates are present on anemophilous pollen, they are more likely to be much smaller than their entomophilous counterparts. Despite protective structures pollen can still rupture due to mechanical stress as well as elevated humidity (Pacini and Hesse 2005).

Atmospheric lifetimes of pollen are of interest, as pollen exposure from the air is critical to triggering an allergenic response (Rapiejko et al. 2007). The release of anemophilous pollen is highly dependent on both temperature as well as humidity (Kuparinen et al. 2009; Sato and Peet 2005; Grote et al. 2003). Wind-born pollen generally travels distances of hundreds of meters or more, well past adjacent maternal neighbors, with one study showing a minimum of 62% of viable *C. longifolium* pollen traveling 200+ meters (Kuparinen 2006; Stacy et al. 2002). Episodic scenarios also occur with pollen traveling hundreds of kilometers over a several-day period. *Betula* pollen has been seen to travel over 2000 km in some instances, over the course of up to 50 hours, and are still able to elicit allergenic reactions (Hjelmroos 1991). In certain circumstances, entomophilous pollen can also be ejected into the atmosphere due to high winds, forming a larger fraction of the pollen load than calm conditions (Dua and Shivpuri 1962), though it is less common. Settling coefficients of pollen impact its ability to be transported large distances or stay suspended in the air. Pollen diameter, morphology, and other properties such as hygroscopicity also directly affect the settling parameters of pollen (Aylor 1975).

Clouds of boreal tree pollen have been observed by polarization LIDAR measurements in Alaska, lofted up to 2 km in the atmosphere from adjacent forests (Sassen 2008). Pollen is frequently detected by LIDAR measurements in Fairbanks during the summer months. Forests are a large contributor of atmospheric bioaerosols, and northern hemisphere boreal forests (specifically the species *Pinus taeda*) in the southeastern United States has been shown to contribute 3.3 Tg of pollen over the course of less than 100 days (Williams and Després 2017). *Pinus taeda* exceeded per-plant production of pollen by corn plants dramatically, despite Zea Mays pollen serving as the current source for modeling pollen emissions and transport (Williams and Després 2017).

1.2.2 Pollen Allergies

The primary health concern related to pollen is the propensity of certain species to induce an allergenic response. Though it is possible for entomophilous plants to be allergenic, anemophilous ones tend to drive pollen allergies. This is because entomophilous pollen is much less likely to be lifted into the air, so the risk of contact is greatly reduced. However, some anemophilous species, such as those in the *Pinus* genus, do not illicit allergenic responses (Spieksma 1990), implying that a combination of allergenic potency and availability of contact is important. It has been suggested that the overall morphology and structure of anemophilous pollen types allow for easier access to the allergenic compounds within the pollen itself (Diethart et al. 2007). Since anemophilous plants pollinate during most of the year, usually peaking for trees in the spring, grasses in the summer, and weeds in the fall, they are a large influencer of human health.

Pollen allergies are prevalent among humans, caused by an allergenic response following ingestion or inhalation of microscopic pollen grains (Douwes et al. 2003; Cohen et al. 1979). Allergies cause an immune system reaction that produces Immunoglobulin E (IgE) antibodies, which then travel to cells and trigger the release of histamines. These histamines are the chemicals responsible for the allergenic symptoms of coughing, sneezing, and inflammation. Many allergens present in pollen particles exhibit cross-reactivity, resulting in allergenic responses to many types of pollen allergens since the IgE epitopes are conserved between those allergenic proteins (Wopfner et al. 2005). Different types of pollen have also been shown to contain multiple allergenic proteins (Léonard et al. 2010; Asero et al. 2006; Wopfner et al. 2005). Though the locations of these allergens vary within the structure of the pollen, many of these allergens are water soluble (Vrtala et al. 1993; Staff et al. 1990), allowing for quick dissemination of these proteins in high humidity or aqueous conditions. Different types of pollen have also been observed to rupture due to high humidity, releasing respirable fragments that further expose allergens deeper into airways and lungs (Taylor et al. 2004; Grote et al. 2003).

Direct inhalation or contact is not the only pathway for the induction of the allergic resonse. Pollutants in the atmosphere can compound these problems in multiple ways. Previous studies have shown that nitration of proteins by atmospheric pollutants, such as nitrogen dioxide and ozone, in already allergenic pollen can lead to an increased allergenic response (Karle et al. 2012; Gruijthuijsen et al. 2006; Franze et al. 2005). Increases in carbon dioxide concentrations have also been seen to cause an increase in the production and growth of allergenic pollen, as seen in *Ambrosia artemisiifolia*, or common ragweed, pollen (Wayne et al. 2002; Ziska and Caulfield 2002). There has been an increase in overall plant colonization in regions in Europe due to increases in carbon dioxide concentrations and human activity (D'Amato et al. 2007). Some pollen species, such as Mugwort pollen, have been seen not to cause allergic reactions in humans unless the pollen was previously contaminated with endotoxins from certain bacteria (Oteros 2019). Oral allergy syndrome is also caused by pollen, in which previous exposure to pollination allergens, or pollen presently existing on the food, causes an allergic reaction from inhalation while eating (Balková 2015; Katelaris 2010).

1.3 Pollen Sampling and Monitoring

1.3.1 Traditional Pollen Sampling

Current atmospheric pollen detection typically relies on a combination of a collection and subsequent analysis technique. The collection mechanisms frequently involve the capture of pollen particles with a sticky grease material, which can do this over time. The Hirst sampler, for example, collects within a spinning drum that slowly spins over the course of a period of time, usually a week (Mullins and Emberlin 1997). The pollen samples collected with this mechanism are then collected and examined by visual microscopy by a palynologist. Collected pollen needs to be prepared prior to microscopy is performed, usually to stain the pollen exine and nothing else. Pollen can then be identified to a taxonomic level, occasionally to the species, by applying an analysis of pollen grains for grain number, size, shape, surface structures, and internal structural details (Weber 2010). These pollen particles need to be identified individually based on the collections over approximately a week, for example, which can be both costly and time-consuming to perform because measurements are typically carried out by trained professionals. Studies of computational image analysis algorithms have shown that many human analysts may be under-performing when compared to emerging algorithm technologies (Mander et al. 2014). Human analysts were seen to produce mean accuracy results of 46.67% to 87.5% when attempting to identify grass pollen species alone (Mander et al. 2014).

1.3.2 Current Monitoring Networks

Monitoring networks for pollen exist around the world, though sampling sites are often widely separated and inefficient. Many countries or continental regions operate national networks of pollen monitors for the purposes of public health information. In continental Europe, for example, a well-developed network of sites (>525) collect data about relative levels of key allergenic pollen and fungal spore species on a daily basis, whereas a smaller network of ~150 stations is operated in the United States (Buters et al. 2018)**.** An interactive pollen measurement global map assembled by the Center for Allergies and Environment in Munich, Germany shows high density of pollen measurement in most of Europe, the Eastern United States, and Japan, additional points around China and Australia, and then at best sparsely scattered measurements across the rest of the world (Buters et al. 2018).

Current pollen identification processes are costly due to the requirement of using technicians trained in the specialized biological identification process, provides data at relatively low time resolution (min. 2+ hr), and leads to poor spatial resolution of sampling sites. For example, only ca. 20 monitoring sites are operated in the vast geographic region of the western United States, and many states have no sites at all (National Allergy Bureau, 2019). Pollen counts at locations between measurement sites are interpolated, and thus the quality of prediction at the local level varies significantly. Local and regional topological effects influence pollen measurement and prediction accuracy for interpolated forecasting (Tseng and Kawashima 2019)**.** Data from pollen monitoring stations is combined with meteorological conditions as input parameters for predictive models. The result is to forecast e.g. the beginning date of pollination season and the relative concentration of key allergenic pollen classes as a function of geography (Pauling et al. 2012; Stach et al. 2008; Galán et al. 2001)**.** These forecasts are frequently then transmitted to the public via news channels and smartphone applications.

1.4 Fluorescence Spectroscopy

1.4.1 Introduction to Fluorescence

Fluorescence spectroscopy can be a useful tool to probe chemical compositions of substances. Electrons present in the matter absorb electromagnetic radiation at varying energy, depending on overall chemical structure. Fluorescence is generally referred to as the ability to absorb electromagnetic radiation in the UV and visible range on the electromagnetic spectrum and re-emit visible light. Absorption of photons by the substance's electrons causes them to be excited, taking them from the ground state to a higher energy level. The energy then goes through internal conversions based on the chemical composition of the molecule, and when relaxing back to the ground state emits a new, lower energy photon as visible light (Atkins and De Paula 1989). This can be seen in the example Jablonski diagram in Figure 1.1. Fluorescence instrumentation has been developed to take advantage of this property. A sample, liquid or solid, is illuminated, frequently with near/deep UV or blue-end visible light, and the subsequent fluorescence is collected, often at a 90° angle. The fluorescent light is refracted through a monochromator or prism to separate the light into individual wavelengths and is measured from that point.

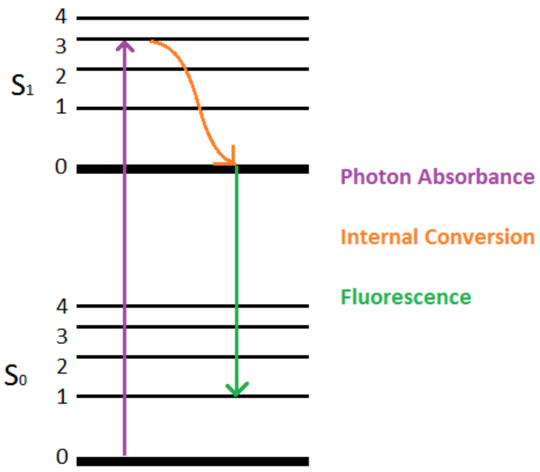


Figure 1.1. Simplified Jablonski diagram representing a typical absorption and re-emission of a photon in the process of fluorescence.

Various instrumentation has been developed to take advantage of fluorescence. Simple fluorescence imaging (Dobrucki and Kubitscheck 2017), adding fluorescent tags (Sahoo 2012), utilizing polarization with fluorescence (Lakowicz 2006), multifocal plane fluorescence microscopy (Prabhat et al. 2004), and even testing the effect of photobleaching as it related to fluorescence are all useful techniques that incorporate fluorescence spectroscopy (Axelrod et al. 1976). Detected fluorescence signals can give insight into chemical composition, or changes in composition, depending on the wavelength, shape, and even ratio of fluorescent signals. In particular, ultra-violet laser-induced fluorescence has become a commonplace technique in the detection of bioaerosols (Huffman et al. 2019; Huffman and Santarpia 2017).

1.4.2 Pollen Fluorescence

Many of the chemical compounds in pollen exhibit auto-fluorescence, the re-emission of light from natural structures, allowing for different fluorescence emission signals to be detected based on the usage of excitation sources (Pöhlker et al. 2013; O'Connor et al. 2011). Chemical structures such as chlorophyll, NADH, proteins, and the pollenkitt on the surface each exhibit different fluorescence characteristics that allow for differentiation in these particles. Phenolics, carotenoids, proteins, chlorophyll *a*, and other biological compounds all exhibit fluorescence modes at differing excitations and emission intensities. In particular, riboflavin, NADH, and tryptophan/tyrosine are three primary biological fluorophores that show large emission signals in the visible range, and are commonly present in most biological materials (Pöhlker et al. 2012).

Fluorescence spectral characteristics of bulk pollen powder have been comprehensively analyzed, frequently presented in excitation emission matrices (EEMs) for individual pollen species (Pöhlker et al. 2013; Hill et al. 2009; Wlodarski et al. 2006; Satterwhite 1990). Based on spectral trends and general molecular composition assignments summarized by Pöhlker et al. (2013), the assessment of spectra analyzed here were grouped into eight spectral regions according to approximate location of spectral peaks: (0) e.g. protein signals; λ_{Ex} 280 nm, λ_{Em} 350 nm (I) e.g. phenolics; λ_{Ex} 280 nm, λ_{Em} 450 nm (II) e.g. phenolics; λ_{Ex} 360 nm, λ_{Em} 450 nm (III) e.g. phenolics; λ_{Ex} 405 nm, λ_{Em} 450 nm (IV) e.g. carotenoids; λ_{Ex} 360 nm, λ_{Em} 500-520 nm (V) e.g. carotenoids; λ_{Ex} 405 nm, λ_{Em} 500-520 nm (VI) e.g. carotenoids; λ_{Ex} 450 nm, λ_{Em} 520-550 nm (VII) e.g. chlorophyll *a*; λ_{Ex} 405 nm, λ_{Em} 675 nm (VIII) e.g. chlorophyll *a*; λ_{Ex} 450 nm, λ_{Em} 675 nm. The first signal is listed as (0) as the wavelength of emission is too low to be seen by this instrument.

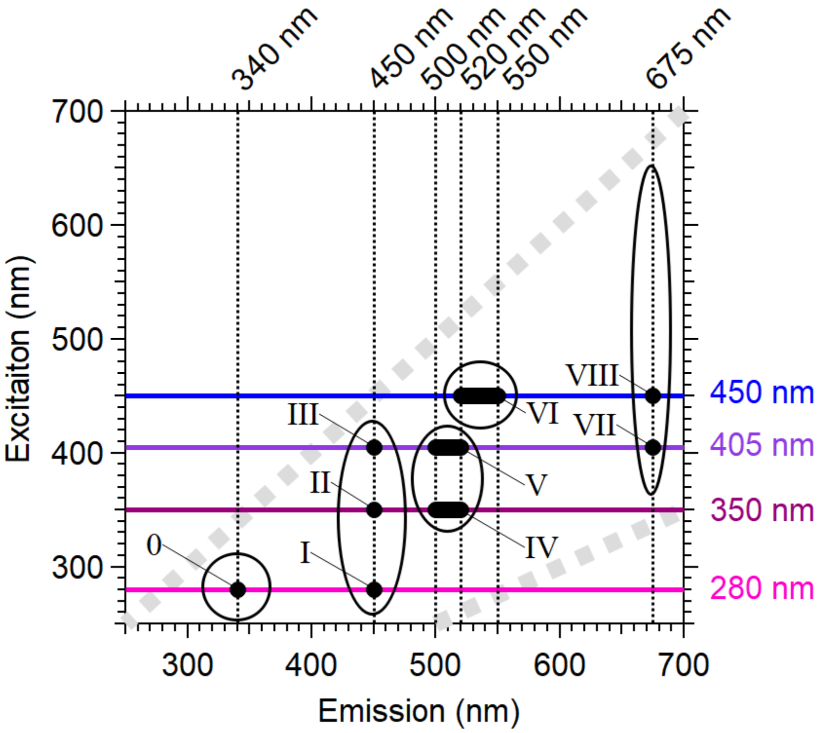


Figure 1.2. An excitation-emission matrix spanning fluorescence modes of pollen from previous studies, as well as the information compiled in this thesis. The four colored lines represent excitation sources available in this instrument, and fluorophore modes above 0 are seen by this instrument. (Adapted from Pöhlker et al., 2013)

1.5 Emerging Techniques for Pollen Analysis

As a result of the challenges of relying on manual identification processes, significant effort has gone into finding automated solutions to replace or supplement existing detection strategies (e.g. Huffman et al. 2019; Wu et al. 2018; Kawashima et al. 2017; Tello-Mijares and Flores 2016; Oteros et al. 2015; Kiselev et al. 2013; Dell'Anna 2010; Allen et al. 2008; Ranzato et al. 2007; Chen et al. 2006). Efforts to automate pollen analysis continue to face technical challenges, and so at present only a few groups have experimented with deploying prototypes of automated techniques (Buters et al. 2018).

The allergenic pollen burden of many regions of Japan is heavily dominated by a single pollen species (Japanese cedar), and so a single-particle light-scattering instrument (KH-3000, Yamatronics; Japan) was developed largely to quantify this pollen type (Kawashima et al. 2007, 2017). The instrument is now functional in networks around Japan (Miki et al. 2019; Kawashima et al. 2007, 2017). The BAA500 (Hund-Wetzlar; Wetzlar, Germany) was develop to mimic the operational process of collection and microscopy analysis, and is being used in small numbers in the ePIN pollen monitoring network in southern Germany (Oteros et al. 2015).

Many examples of ultra-violet laser-induced fluorescence (UV-LIF) instruments have been utilized to selectively detect biological fluorophores in atmospheric particulate matter and have been applied not only for pollen detection, but for rapid detection and classification of a wider range of biological aerosol types including bacteria and fungal spores (Huffman et al. 2019; Fennelly et al. 2017; Pöhlker et al. 2012, 2013; Després et al. 2012; Hill et al. 2009). Data from a new instrument (Swisens AG; Horw, Switzerland) using real-time holography measurements was shown to be applied to convolutional neural networking systems to identify pollen at a taxonomic level (Sauvageat et al. 2019). The Wideband Integrated Bioaerosol Sensor (WIBS, Droplet Measurement Technologies; Longmont, Colorado), for example, uses two excitation sources (280 nm and 370 nm) to selectively target biofluorophores, capturing fluorescence signal with coarsely binned resolution of two channels per emission spectrum (Savage et al. 2017; Hernandez et al. 2016; Gabey et al. 2010; Kaye et al. 2005). The WIBS has been applied for pollen detection (Ruske et al. 2018; Calvo et al. 2018; Savage et al. 2017; Perring et al. 2015; D. O’Connor et al. 2014a; Healy et al. 2012), but with only limited success (Savage and Huffman 2018; D. O’Connor et al. 2014a; Kiselev et al. 2013). The Rapid-E (Plair SA; Geneva, Switzerland) acquires fluorescence spectra in 32 channels after excitation by a 400 nm laser and also records fluorescence lifetime and time-resolved light scattering signal in order to more finely differentiate between pollen species (Kiselev et al. 2011, 2013). The Rapid-E has been applied to ambient pollen monitoring in several studies, and shows the ability to discriminate between certain groups of pollen types with roughly 90% accuracy (Šaulienė et al. 2019; Crouzy et al. 2016).

While these few examples of instrumentation able to identify or differentiate broad classes of pollen are under investigation for application for monitoring networks, their purchase cost is high at tens to hundreds of thousands of dollars per unit. As a result, the need exists both to improve upon recognition capabilities and to dramatically reduce the purchase cost of pollen sensors.

Within the last few years, a separate paradigm of pollen detection has also become popular, shifting toward smaller, relatively inexpensive instrumentation. In most of these cases, the physical principles of detection are based on light-scattering, pattern recognition, or holography, using more advanced analysis computing to differentiate pollen types using field-portable instrumentation. One such prototype sensor generates diffraction holograms associated with individual particles, and deep learning techniques are then utilized to process and subsequently classify, or label, the measured particles from the hologram (Wu et al. 2018)**.** The sensor was shown to successfully separate a mixture that included three species of pollen (Bermuda grass, oak, and ragweed), two fungal spore types, and common dust, with a classification accuracy of 94% (Wu et al. 2018). Another recently available commercial sensor is the Pollen Sense™ (Pollen Sense, Salt Lake City, Utah), which is a portable and relatively low cost (~$8,000) sensor that utilizes a combination of visual microscopy and image analysis techniques to identify pollen types as well as other large particles (i.e. ~5 μm)

1.6 Research Aim

Detection and classification of pollen populations is essential to efficiently model and forecast the allergens carried by these particles. Currently existing monitoring networks exist but consist of costly and time-consuming techniques that do not encourage widespread utilization. This results in poor spatial coverage, leading to potentially inaccurate forecasts. In the United States, stations monitoring pollen are particularly sparse, and a system of automated sensors contributing to the network of available pollen information would be extremely beneficial.

Emerging technologies in UV-LIF, holography, and other microscopic techniques, have shown promise in the detection and classification of pollen, and bioaerosols in general. Still, the majority of the techniques that have proven successful have not seen commercialization at a price-point that allows for wide distribution. We have developed a new single-particle fluorescence spectrometer that allows for the detection and subsequent classification of pollen particles. This sensor is comparatively inexpensive (e.g. <$5000 current prototype fabrication cost) and can detect and classify different types of pollen with relatively high accuracy (>90%) in most cases, as will be discussed. The ability to detect and classify certain subsets of local, allergenic pollen from background materials, including other pollen, will represent a leap forward in alerting the public to the potential of an allergenic response.

This thesis demonstrates the achievement of the following research goals:

- To show early development of an inexpensive, single-particle fluorescence spectrometer that is capable of high-resolution fluorescence spectroscopy, as well as the operation from collection to analysis of individual particles (Chapter 2).
- To suggest improvements on instrumental design for focus on pollen detection and classification, to demonstrate the spectral range obtained by the instrument, and also suggest strategies for instrumental calibration. (Chapter 3).
- To introduce clustering and classification strategies that were utilized with the fluorescence/size data obtained from the instrument (Chapter 4).
- To apply these strategies to pollen samples collected from local plants, as well as discuss spectral anomalies and various factors such as collection time, sample age, and other characteristics (Chapter 5).

The instrument presented in this thesis is also patented under US Patent S20160320306A1 (Huffman and Huffman 2019)

Chapter Two: Design and Basic Instrumental Operation of a Newly Developed Single-Particle Fluorescence Spectrometer

2.1 Introduction to Fluorescence

Fluorescence spectroscopy is a useful tool to probe chemical compositions of substances. Electrons present in the matter absorb electromagnetic radiation at varying energy, depending on overall chemical structure. Fluorescence is generally referred to as the ability to absorb electromagnetic radiation and re-emit light in photons with less energy. Absorption of photons by the substance's electrons causes them to be excited, taking them from the ground state to a higher energy level. The energy then goes through internal conversions based on the chemical composition of the molecule, and when relaxing back to the ground state emits a new, lower energy photon as visible light (Atkins and De Paula 1989). Fluorescence instrumentation has been developed to take advantage of this property. A sample, liquid or solid, is illuminated with an excitation source, and the subsequent fluorescence is collected, often at a 90° angle. The fluorescent light is refracted through a monochromator or prism to separate the light into individual wavelengths and is measured from that point.

Various instrumentation has been developed to take advantage of fluorescence. Simple fluorescence imaging (Dobrucki and Kubitscheck 2017), adding fluorescent tags (Sahoo 2012), utilizing polarization with fluorescence (Lakowicz 2006), multifocal plane fluorescence microscopy (Prabhat et al. 2004), and even testing the effect of photobleaching as it related to fluorescence are all useful techniques that incorporate fluorescence spectroscopy (Axelrod et al. 1976). Detected fluorescence signals can give insight into chemical composition, or changes in composition, depending on the wavelength, shape, and even ratio of fluorescent signals. In particular, ultra-violet laser-induced fluorescence has become a commonplace technique in the detection of bioaerosols (Huffman and Santarpia 2017).

2.2 Bioaerosol Fluorescence and Instrumentation

Biological particles contain a mixture of molecular components that can be probed to differentiate them from abiological material. Some of these molecules possess intrinsic fluorescence (autofluorescence) properties that can be exploited to spectroscopically detect and characterize PBAP (Pöhlker et al. 2012). As a result, UV-LIF technologies have been developed for widely different applications in a number of industries and research fields (Kiselev et al. 2011; Kaye et al. 2005; S Hill et al. 1999; Hairston et al. 1997). The growing number of commercially available UV-LIF bioaerosol sensors traditionally require high upfront purchase cost (ca. $100k or more) and relatively skilled operators to interpret complex environmental data (Huffman and Santarpia 2017; Sodeau and O’Connor 2016; Crawford et al. 2015; Robinson et al. 2013). These technologies typically offer excellent time resolution (seconds to minutes), however they frequently also suffer from comparatively weak ability to discriminate between particle types due to poor spectral resolution. The majority of commercial UV-LIF instruments are limited to either one or two wavelengths of excitation and integrate emission intensity into 1-3 total channels, which allows only a limited set of fluorophores to be probed and limits quality of discrimination between particle types. For example, the wideband integrated bioaerosol sensor (WIBS; Droplet Measurement Technologies) utilizes two excitation sources (280 nm and 370 nm), chosen to excite commonly occurring biofluorophores such as tyrosine and NADH, respectively (Pöhlker et al. 2012; Kaye et al. 2005)

Many other fluorophores present in both biological and abiological material can emit fluorescence in overlapping wavebands, thus, resulting analytical selectivity and discrimination between particle types can be poor due to limited number of channels of both excitation and emission (Savage et al. 2017). Recently, a new generation of UV-LIF instruments have become commercially available to interrogate bioaerosols at higher spectral resolution. Instruments like the BioScout (Environics, Ltd.) (Saari et al. 2014), Rapid-E (Plair) (Kiselev et al. 2013), Multiparameter Bioaerosol Spectrometer or MBS (University of Hertfordshire) (Ruske et al. 2017), and the Spectral Intensity Bioaerosol Sensor or SIBS (Droplet Measurement Technologies) deliver fluorescence spectra recorded with 8-32 channels of resolution (Könemann et al. 2019b; Huffman and Santarpia 2017). These instruments generally cost as much or more, however, than earlier generation instruments with lower spectral resolution and require even more expertise to interpret spectra.

To differentiate between particulate analytes of interest and interfering species often requires the top end of UV-LIF instrumentation or innovative computing analysis strategies (Robinson et al. 2013; S Hill et al. 1999). For example, one UV-LIF instrument, the Single-Particle Fluorescence Spectrometer, developed in part by the Army Research Laboratory provides high spectral resolution for each of several excitation wavelengths, though the cost and complexity of the instrument would prevent widespread commercialization or consumer application (Pan et al. 2007; S Hill et al. 1999). In addition to instruments that have been developed to utilize laser excitation sources at wavelengths similar to those chosen for the WIBS (i.e. ~280 nm or ~360 nm), many recent instruments have employed use of the 405 nm diode laser, which was made inexpensive by its application to Blu-Ray^{TM} video technologies. Availability of relatively inexpensive UV light emitting diode (LED) technology is also becoming increasingly important for PBAP detection, as will be discussed with respect to the instrument introduced here (Zhang et al. 2013).

There is growing interest to monitor several classes of PBAP, such as pathogenic or allergenic fungal spores and pollen, in home and occupational health settings (Fröhlich-Nowoisky et al. 2016; Douwes et al. 2003). Significant effort in recent years has been focused on the development of automatic techniques for pollen counting, some of these utilizing UV-LIF technologies. These techniques are very expensive and have yet to find wide-scale application (D. O’Connor et al. 2014a; Kiselev et al. 2013; Pan et al. 2011; Kawashima et al. 2007; Rodriguez-Damian et al. 2006; Chen et al. 2006; Aronne et al. 2001). For these reasons we have developed a simple technique to characterize individual particles, approximately micron-sized or greater, by their fluorescence spectra, achieved at much lower instrument purchase cost and with much improved spectral resolution and ability to discriminate between particle types than widely available UV-LIF aerosol sensors.

Among the purposes of developing this instrument are to complement current instruments by improving spectral discrimination of particles at a significantly reduced cost and complexity in order to enable wider-scale application for research and monitoring. The instrument was also designed as a tool for the investigation of particles and for education about fluorescence spectra by citizen scientists or schools of various levels. Described here is an instrument that simultaneously provides fluorescence or scattering spectra of many individual particles collected onto a substrate.

2.3 Instrumental Design

The inception and initial work on the developed fluorescence instrument is described in Huffman et al. 2016. The instrument described in this thesis is functionally similar to a classical spectroscope, though innovative in its usage on micron-sized particles. Fluorescent light is projected through simple microscope optics, which is then dispersed through a transmission grating to split the light into individual components. Fluorescent signals can then be observed visually through the eyepiece that focuses the dispersed light. Normal spectroscopes utilize a light entrance slit, while this instrument examines small particles that act as point sources in place of a slit. Astronomical spectral studies have used slitless spectroscopy since the late 1800s, where Edward Pickering constructed a new method to image several stars in a single image by including a prism before the photographic plate (Bunch and Hellemans 2004). Slitless spectroscopy is still used to analyze sparsely populated star fields and gain emission spectra information (Sachkov et al. 2014; Kümmel et al. 2009; Stanghellini et al. 2002). Micron-sized particles can be viewed similarly to very large, distant objects like stars. These astronomical applications were the inspiration of the instrument presented in this thesis.

There are other applications that utilize a similar framework, though this is the first instance of this technique being applied to atmospheric aerosol particle analysis (Xiong et al. 2013; Cheng et al. 2010; Lakowicz 2006). Since aerosol particles do not autoluminesce and so an external excitation source must be applied. Optical spectroscopy that results in the emission or scattering of visible light can be used in this instrument. An example of elastic versus inelastic scattering comparisons is shown in the diagram for Figure 2.1, where particles deposited onto a substrate, such as a glass slide, are analyzed via a tungsten lamp or fluorescence excitation source and observed visually. These resultant streaks appear visually similar to emission signals from astronomical objects previously described.

The instrument utilizes a dissected microscope to magnify the emission signals. The particles are deposited onto a glass slide, and this slide is placed into a microscope X-Y positioner for easy translation of particles. The particles are illuminated by a simple tungsten lamp (General Electric, Miniature Lamp 210, B6, 6.5 V**)**, four independent excitation sources (450 nm laser, ThorLabs CPS450, Newton NJ, 405 nm laser, Power Technology Incorporated 9-0407-A560-0-0, Little Rock AR; 350 nm LED, QPhotonics, UVCLEAN350-5; 280 nm LED, QPhotonics UVTOP280, Ann Arbor MI), or a Helium-

Neon laser at 632.8 nm (Meredith Instruments HNS-LL-1, Peoria AZ). The light, either refracted or emission from fluorescence, is magnified through a simple student microscope optical setup (Model 656/98, SWIFT Microscopy, Carlsbad, CA). The light is then dispersed using a transmission grating (300 grooves mm^{-1}; ThorLabs, Inc., GT25-03) located on the pivoting point of an optical rail. At the end of the optical rail, two CCD cameras have been used as inexpensive detectors: A color camera (Canon Powershot A2300 HD, Canon Inc., Tokyo JP), and a monochrome camera (Lumenera Infinity 2-1R, Lumenera Corporation, Ontario CN). This typically results, at 10x magnification, in an approximate area of 1.0 mm wide by 1.0 mm high being the size of a typical sample window. In the cases of fluorescence emission, the excitation source is blocked out using a requisite long-pass filter from Edmund Optics (No filter for 280 nm; 435 nm filter for the 350/405 nm excitations; and the 470 nm filter for the 450 nm excitation from the Edmund Optics #832916-10 filter kit).

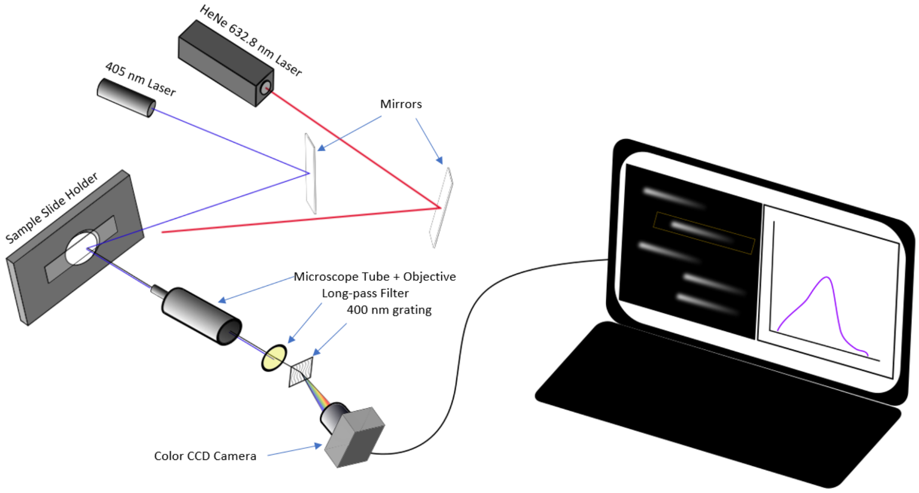
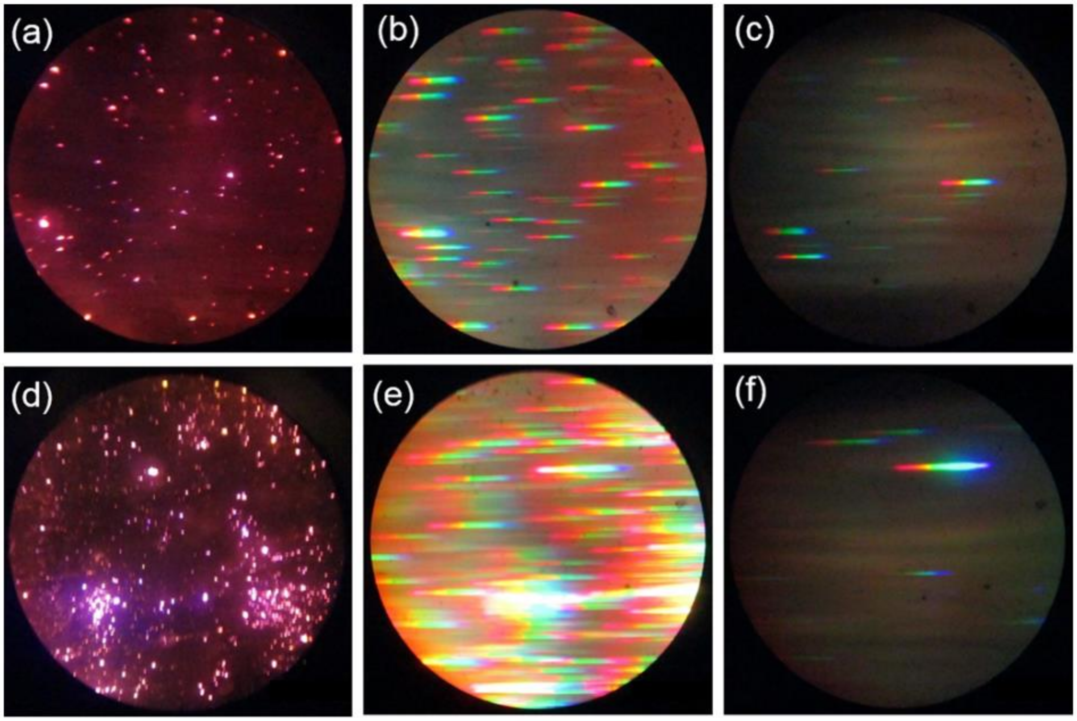


Figure 2.1. Current iteration of the desktop instrument described in chapter 2 as well as the updates from chapter 3. Details in chapter 2 related directly to the 405 nm source, red laser, and the original color camera sensor.

2.4 Instrumental Operation

Spectral collection is performed in many distinct steps. The sample of particles, present on a glass slide, is illuminated by an excitation source. Light is scattered, elastically (not absorbed) or inelastically (absorbed and re-emitted) as a function of the chemical components within the particles. Elastic scattering is highly dependent on the contact angle and structure of what is being illuminated, while fluorescence is emitted isotropically. Some of this light is directed towards the objective lens, which is collected and magnified in the microscope optics. For inelastic scattering, the excitation source is filtered out through a long-pass optical filter. All light is dispersed through the transmission grating, which then is collected with a digital camera. For elastic scattering, this can be seen in Figure 2.2 (Figure reproduced from Huffman et al., 2016) for an ambient collection of particles (top row) and ground quartz particles (bottom). Images were taken with a Canon Powershot A2300. The first column in Figure 2.2 represents a simple micrograph taken with the camera at the 0th angle with respect to the grating position, with no grating in place. The second column in the Figure shows the elastic scattering of a tungsten bulb from the particles, taken at 8° with respect to the grating, allowing for the elastic light to be dispersed into the individual components of light. The last column of the Figure represents the camera being in the same position, but a 405 nm excitation source illuminating the particles instead of a white light source, and the excitation source filtered out. The difference in columns two and three show differences in particle number between fluorescence and non-fluorescent particle types. In this example, ~2% (6 of 200-250 fluorescent) of the total number of quartz particles are fluorescent, likely due to contamination from handling these samples in a relatively dirty lab space. In contrast, ~7% (7 of 49 fluorescent) of the ambient outdoor sample are fluorescent particles.



.

Figure 2.2 Differences in scattering and fluorescence images for two samples. The top row (a-c) represent an ambient sample collected via ambient deposition and the bottom row (d-f) represent ground quartz particles introduced via mechanical deposition. Column 1 (a,d) shows a simple micrograph from the instrument, column 2 (b,e) shows the elastic scattering using a tungsten light, and column 3 (c,f) shows the corresponding fluorescence of these particles with the 405 nm laser diode (Figure reproduced from Huffman et al., 2016)

For samples that represent pure bioparticle standards, the ratio of fluorescent to elastic particles is closer to 100% fluorescence. Figure 2.3 shows the process behind spectral collection, which is conceptually similar to Figure 2.2. Figure 2.3a is a dark field image of paper mulberry pollen particles (*B. papyrifera*; 12–13 µm; Thermo Fisher Scientific).

Figure 2.3b shows the same set of particles illuminated with both the 632.8 HeNe laser and the 405 nm excitation laser, without the excitation being blocked. This type of image is the primary calibration image used in our analyses and is used in the two-point calibration described in the next section. Figure 2.3c shows the same particles illuminated by the white-light tungsten bulb, elastic scattering. Figure 2.3d is fluorescence emission after illumination by the 405 nm laser, with the 405 nm spot filtered out with the 435 nm blocking filter.

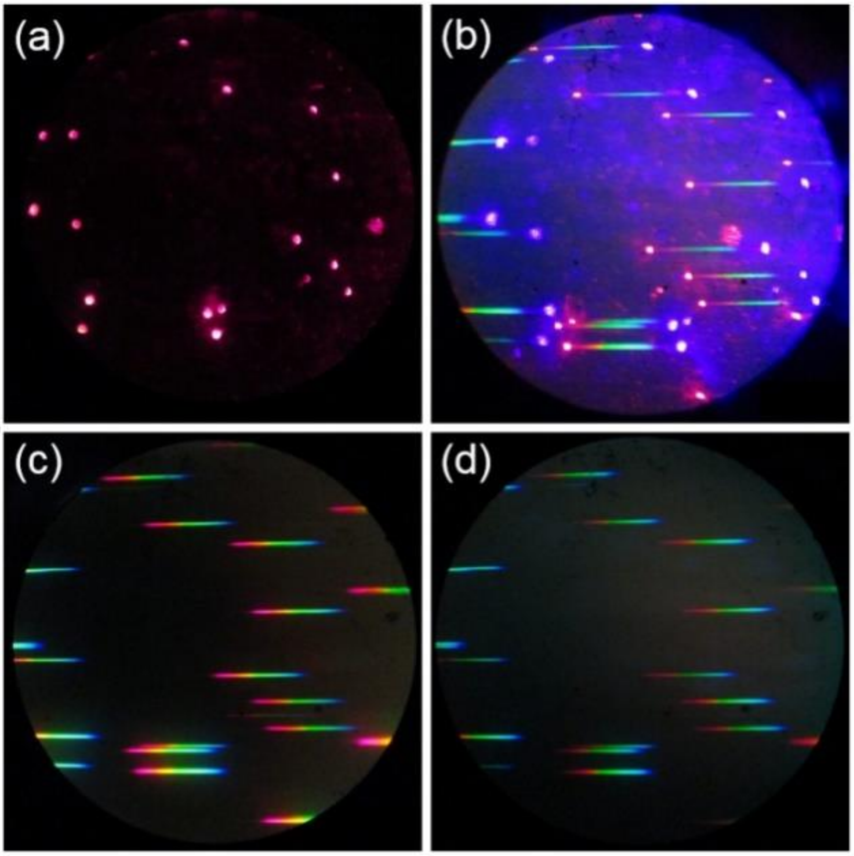


Figure 2.3 Progression of spectral collection of paper mulberry pollen particles collected on a slide under 100x magnification. (a) A microscopy image with illumination from the HeNe Red laser. (b) Illumination with the 405 nm and HeNe laser. (c) white light tungsten bulb scattering. (d) fluorescence using the 405 nm excitation. (Figure reproduced from Huffman et al., 2016)

Particles are analyzed using the same camera every time, though it is worth noting that digital cameras include settings that allow for variable capture options like gain or exposure time. This is a useful tool, considering fluorescence scales with many factors including size, quantum yield, or fluorophore concentration, because it enables capture of spectra for a wide range of particles. When collecting fluorescence spectra, care is taken to get a reasonable image that is above the limit of quantitation (LOQ), but also below the saturation limit of the detector itself.

2.5 Analysis of Collected Spectra

The spectral signals collected by the instrument need processing prior to interpretation. The open source image analysis software, ImageJ, is utilized to pull the raw numerical information out of each spectral image. The images are imported into the program, and a "region of interest" (ROI) is drawn encompassing the entirety of the spectral signal in length, and high enough to cover the height of the particle. A profile is taken from the ROI of the particle, averaging the mean grey value in the Y dimension of the ROI, and reporting this along the X value of the ROI.

Figure 2.4b represents the image utilized for calibrating these spectral signals. Though there is a dispersion swath of fluorescence seen here, the important features of this image are the red and blue laser scattering points present, which appear as larger, washed out dots on each end of the swaths. These dots represent the 632.8 and 403.5 nm laser scattering and can subsequently be translated from pixels to nanometers utilizing a simple ratio calculation. This needs to be done on a particle by particle basis, as the inexpensive optical components may lead to differences in that ratio due to chromatic aberration changes across the lens. As such, the calibration images are collected for every single image, and the calibration multiplier from the following equation is utilized to calibrate:

$$\underline{\text{Eq 2.1.}}\ \mathrm{M} = \frac{(\text{Red Laser } \lambda - \text{Blue Laser } \lambda)}{(\text{Red Laser Pixel Center} - \text{Blue Laser Pixel Center})}$$

For each spectrum, the blue laser pixel center is subtracted from the overall X axis, moving that point to zero. The X axis is then multiplied by the "M" multiplier in the above equation, which changes for each individual particle, and will also change depending on the image magnification or optical rail angle of the camera. The true wavelength value of the blue laser, 403.5, is then added to the X axis of the spectra. The resulting spectral output can be seen in Figure 2.5, showing three output spectra from the above paper mulberry particles shown.

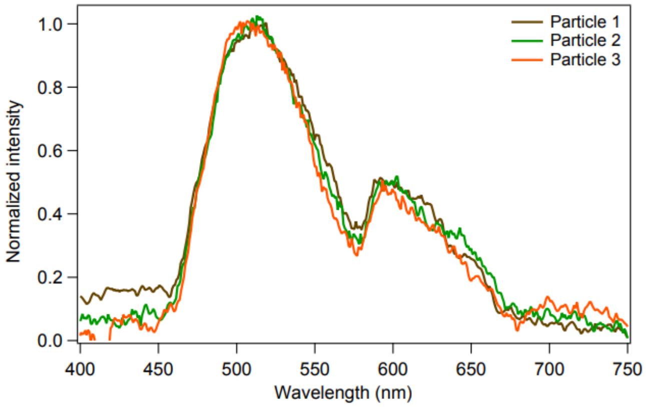


Figure 2.4. Resultant normalized spectra from three paper mulberry pollen particles from the above Figure 2.6d, taken by a color CCD camera (Figured reproduced from Huffman et al., 2016)

2.6 Sample Collection Process

As previously mentioned, the particles analyzed by this instrument are first deposited onto an optical slide. Deposition is achieved differently for various types of particles. The overall goal of any time of deposition is two-fold: (1) producing a monodispersed layer and (2) producing layers that are dispersed enough to measured multiple particles simultaneous, though not such high density that spectral signals overlap. Simple solid phase particles that were standards were introduced onto the optical slide via mechanical deposition. Polystyrene Latex Spheres (PSLs), small uniform plastic spheres, were used widely in the development and characterization of the instrument.

Larger PSL sizes (>8.0 μm) were also deposited with this method, though these were often present in a liquid phase requiring dilution, depending on the solution density, and then dropped onto the slide. These were desiccated over the course of 12 hours to ensure full evaporation. However, this method frequently resulted in lines of dried PSLs along the edge of the evaporation line, leading to large areas of unusable sample. Smaller PSLs (<8.0 μm) utilized an in-house aerosolization mechanism, in which the PSLs were diluted and introduced into aspirators. These were then pushed through the system in diluted samples and pulled into an in-house developed 3D printed impactor to sample using optical slides. Smaller pollen samples were deposited with this method, though even appreciably low sampling times led to very saturated samples, making spectral collections difficult.

2.7 Conclusion

Many bioaerosol analysis tools have been developed or commercialized, though tend to be extremely expensive. Described in this chapter was the initial iteration of the instrument and the primary method of operation involving both elastic and inelastic scattering of deposited particles onto a substrate. Using scattering differences, quick information about fluorescent particle composition of atmospheric samples can be gained by examining the differences in these scattering profiles. It is critical to examine the various parameters of the instrument, and work towards a miniaturized, inexpensive platform that can analyze and classify many pollen particles simultaneously.

Chapter Three: Benchtop Model – Improvements on Design and Analysis

The information present in this chapter has been previously published. The information for the first three Figures was adapted by from Huffman et al. 2016, and the information for the subsequent Figures in the chapter has been discussed separately and published by Swanson and Huffman in 2019.

3.1 Introduction

Building on the general design presented in the previous chapter, application and characterization of the single-particle fluorescence spectrometer is shown here. In this chapter, the implications of a monochrome camera and three new excitation sources are introduced. A variety of instrumental parameters that can affect spectral signals are accounted for to ensure reproducible spectra. These calibration techniques were developed using polystyrene latex spheres, or small uniform plastic beads. An image analysis interpolation technique was also developed to account for differences due to inexpensive or inconsistent optical and excitation defects. These developments push the instrumentation towards a platform capable of detecting and classifying between particle types.

3.2 Design Improvements

3.2.1 Monochrome Camera

The images shown until in Chapter 2 were taken with a simple point-and-shoot Canon Powershot A2300 HD, which comes with several drawbacks. The first issue with these inexpensive color cameras is the Bayer filter over the CCD, which distorts the raw light signal detected by the sensor. This distortion can be seen in Figure 2.8, for example, which seems to show two peaks for each emission spectra. This is due to the three-color pixel filter, blue, green, and red, used to produce color from the total light intensity. The peak sensitivities of light filtered through then correspond to the peaks of these pixel filters. Inexpensive color cameras also frequently possess infrared filters, which cut off spectral signals shortly into the red. Even raw scattering of a blackbody radiator, such as a tungsten bulb, will look fragmented and have multiple peaks. Figure 3.1 this effect, in which the scattering spectra of a salt (NaCl) particle for the color (red trace) and monochrome (black trace) are shown. Each camera, without added optical filter, should exhibit the same response curve. For reference, a blackbody radiatior (3000 K) was calculated and multiplied by the response of the reported CCD sensitivity for the Lumenera 2-1R camera. Seen in Figure 5, it is obvious that the color camera scattering curve is not matching up with the theoretical blackbody curve, and it appears that there are three curve features. This is introduced by the Bayer filter prior to the CCD. Similarly, the spectra appear to end pre-maturely after roughly 660 nm, which is a problem introduced by the infrared filter placed in front of the filter as well.

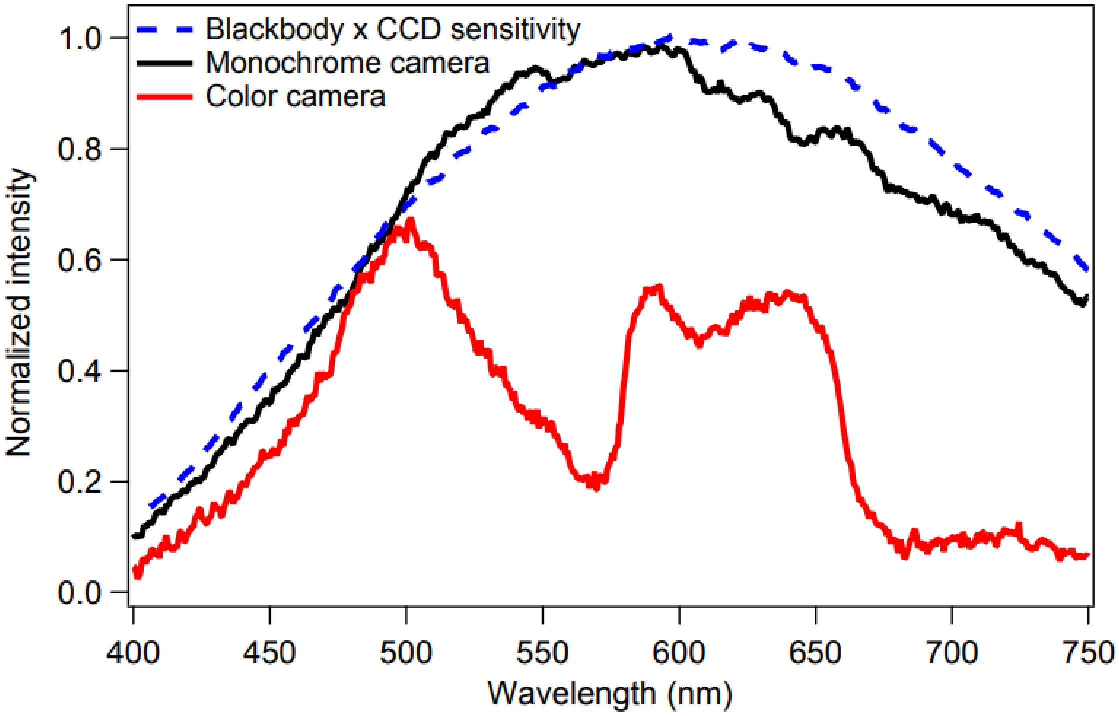


Figure 3.1. Comparison on the scattering of light from a tungsten bulb on an NaCl particle for the color camera (red), the monochrome camera (black), and for the calculated blackbody radiator curve at 3000 K multiplied by the sensitivity of the CCD in the Lumenera camera. The blackbody and monochrome curve were normalized to 1.0, though the red curve was normalized arbitrarily to fit in the shape of the blackbody curve.

Fluroescence singals show these pixel biases as well, displayed in Figure 3.2, which fluorescence signals from an individual, identical Kentucky Bluegrass pollen particle are compared between cameras. Figures 3.2a and 3.2b show the viewable area differences between each type of camera, and the particle of interest is highlighted in the light green box. The resultant signals for each sensor are shown for the color (red) and monochrome (black) cameras in Figure 3.2c against a reference spectrum from a fluorescence microplate reader (Infinite M1000 Pro, Tecan, Mannerdorf, Switzerland) on a black 96-well plate (Fisher Scientific, 07-200-329). The important difference to mention here is that the microplate reader cannot detect the fluorescence on a single-particle basis, so there are discrepancies between the plate reader and either of the camera images. All of the signals were normalized similar to the previous figure. The curves represented here for the monochrome camera signal and the bulk signal are very similar in the main fluorescence curve, when normalized. This changes when looking at secondary peaks, because of variability within individual particles being measured (Pöhlker et al. 2013; Boyain-Goitia et al. 2003). The main structure of the reference curve, however, is clearly conserved. The color camera signal, in contrast, looks extremely different. There appears to be two separate curves, in different maximum positions than either of the other detectors, an extra shoulder on the blue end of the spectra, and the chlorophyll a signal from 675 nm is completely absent due to the infrared filter.

It is important to note that the disadvantages of the color cameras, i.e. the need to calibrate for the color pixel filter as well as the missing 675 nm peak is problematic, even considering the overall cost of instrumentation. It may be possible that a platform utilizing a color camera can adequately detect and classify ambient pollen particles, though introducing more variables into the overall aim of the system to produce a proof-of-concept was not desirable. Still, the color images themselves are much more interesting to look at, and much easier to orient oneself to quickly.

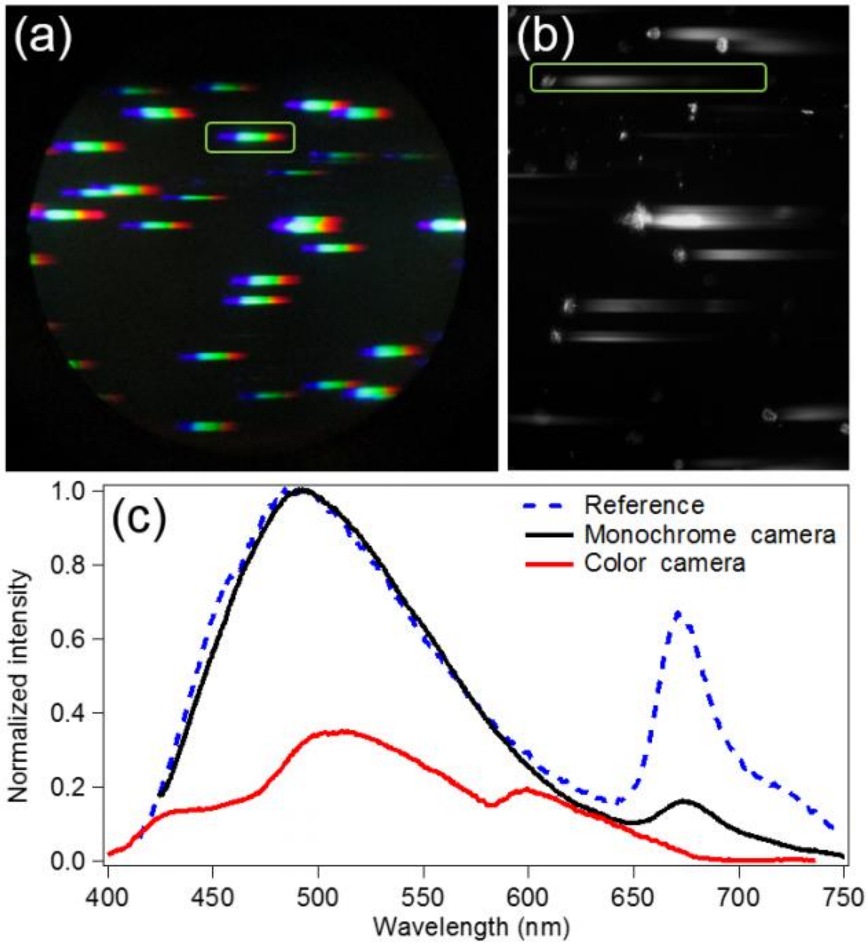


Figure 3.2. Emission signals from a single Kentucky bluegrass (*Poa pratensis*) taken from a color and monochrome camera, excited by a 405 nm diode laser with a 420 nm long-pass blocking filter. The micrograph dispersion images are shown for the color camera in (a), for the monochrome camera in (b), and then the spectral signals are shown in (c) with a reference spectrum from a fluorescence bulk sample of the same particle type.

3.2.2 Multiple Excitation Sources

The original proof-of-concept design implemented fluorescence excitation using a 405 nm laser diode (50 mW; Power Technology, Little Rock, AR). Since that introduction, three additional excitation sources have been added to the instrument. Two UV-LEDs, 280 nm (0.33 mW, with ball lens; QPhotonics, Ann Arbor, MI) and 350 nm (5.0 mW, with flat window and focusing lens tube; QPhotonics), were introduced to the system as well as a 450 nm laser diode (4.5 mW; Thorlabs). The excitation sources were chosen to maximize the breadth of information accessible from individual particles by probing at wavelengths near the excitation maxima for a range of key biofluorophores broadly present in many species of pollen and other PBAP (Fröhlich-Nowoisky et al. 2016; Després et al. 2012). Using all four excitation sources in concert, each particle can be viewed as a total of four emission spectra. This ensemble provides information similar to an excitation emission matrix (EEM) but simplified to the most important excitation wavelengths. A total of two optical filters are employed to block the transmission of elastically scattered light. A 430 nm long-pass filter is used with the 350 nm and 405 nm sources, and a 465 nm long-pass filter is used with the 450 nm source (Edmund Optics, Barrington, NJ). Due to the narrow wavelength distribution emitted from the lasers (405 nm and 450 nm) the filter cut-off can be relatively close to the source wavelength. In contrast, the filter applied after the 370 nm LED requires filter cut-off at much longer wavelength due to the broad range of wavelengths emitted. No optical filter is required after the 280 nm LED source, because the lower wavelength range of detection is limited by the silicon camera CCD (charge-coupled device) whose detection sensitivity drops to zero at ca. 400 nm(SONY n.d.). In an attempt to keep the instrument producible at low cost, standard glass optics are used throughout the optical path, as well as a standard silicone CCD camera. These optical components inherently limit the range of emission spectra detection to ca. 400 - 800 nm. White light scattering spectra are also acquired using an incandescent tungsten filament, with a peak wavelength corresponding to that of a blackbody radiator at ca. 3000 K (Huffman et al. 2016). The white light beam is directed at the microscope stage using a fiber optic light guide (42-347 Edmund Optics; Barrington NJ) 36 inches in length, which was added to reduce the bulkiness of the light source in close vicinity of the microscope stage. Approximately 58% of the light from the original source is projected in a wide cone from the light pipe to illuminate the particles. All optical components are mounted on an optical breadboard placed inside a plywood box, painted black on the interior surface to reduce reflection, with a hinged lid to access the instrument.

As briefly summarized above, images of individual particles can be acquired without filters (Figure 3.1a), showing illumination from both the excitation and HeNe calibration source. Once the filter has been flipped into place, a second image is collected that translates individual particles into streaks with long-axis relative to emission wavelength and short-axis approximately equal to the particle diameter (Figure 3.3b). It is important that many particles be viewable within a single image and under illumination by one excitation source at a time. This allows the simultaneous collection of many emission spectra in a single image at comparatively low-cost, in contrast to confocal fluorescence microscopy, for example. The process described here thus allows the relatively rapid observation of differences between individual particles utilizing a combination of scattering and fluorescence spectra and also enables differentiation between fluorescent from non-fluorescent particles at a glance.

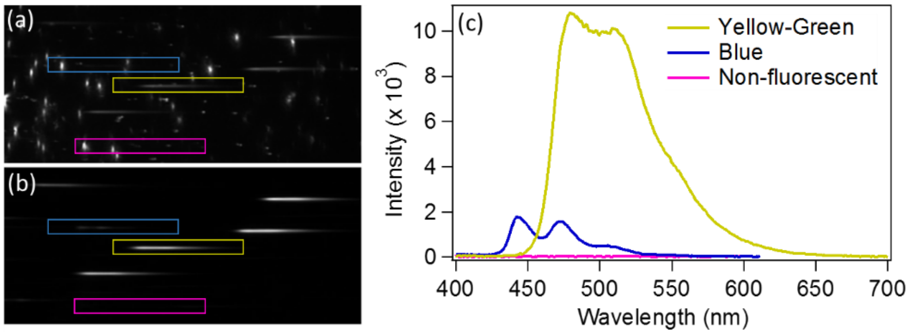


Figure 3.3. Spectrally-dispersed micrographs of a group of many PSL particles, shown (a) without blocking filter (wavelength calibration image) and (b) with blocking filter (fluorescence only). Emission spectra of three specific PSL particles highlighted on micrographs (c), plotted with uncorrected emission intensity. Particle size: Yellow-Green (10.0 µm), Blue (2.0 µm), Non-fluorescent (10.0 µm).

To highlight the ability of the instrument to rapidly differentiate between particle types by investigating spectra, Figure 3.3(c) shows fluorescence spectra extracted from the image in Figure 3.3(b) and acquired from three different types of polystyrene latex spheres (PSL) particles. To produce spectra, boxes are drawn tightly around individual streaks [i.e. Figure 3.3(b)], and detected light intensity is averaged across the y-dimension of the box to achieve a spectrum as a function of location on the CCD. Spectra are then calibrated into wavelength using a twopoint calibration (i.e. Figure 3.3a), as discussed in more detail by Huffman et al. (Huffman et al. 2016). It should be noted that images with a high density of particles can lead to overlapping spectral streaks, which can complicate interpretation of spectra, and so particle collection should be optimized accordingly. Particles deposited close to the edge of the viewable area can also produce spectral streaks that extend beyond the detectable region, reducing the ability to record those spectra. Two PSLs that contain fluorescent dyes excitable at the 405 nm wavelength were used here. The third particle type used for spectral contrast is labeled non-fluorescent by the manufacturer (07310; Polysciences; Warrington PA), though the name is somewhat of a misnomer, because the polystyrene polymer can fluoresce at low quantum yield due to repeating monomer units containing a conjugated aromatic ring (Könemann et al. 2017; Savage et al. 2017). It is important to note that each of the three particle examples have different particle sizes (see Figure 3.3 caption) as well as different fluorescence properties. Based on these differences, the emission spectra for each type of particle are different both in emission intensity and wavelength of peak emission. The ability of the instrument to distinguish these spectral differences is essential to identify and differentiate PBAP of different types.

3.2.3 Range of observable fluorophores

The combination of full-spectra measurements from multiple excitation sources enables a wide range of fluorescence data to be acquired. Six types of PSLs, each commercially doped with different fluorophores, were measured using each excitation source (Figure 3.4). Chosen PSLs varied in dye type as well as particle size. Spectral intensity was normalized to unity for each individual spectrum due to differences in absolute intensity from fluorophore composition and particle size. Normalization was performed here in order to qualitatively highlight the approximate wavelength range of measurable emission spectra by this technique, while acquisition of absolute spectral intensity is discussed in a following section.

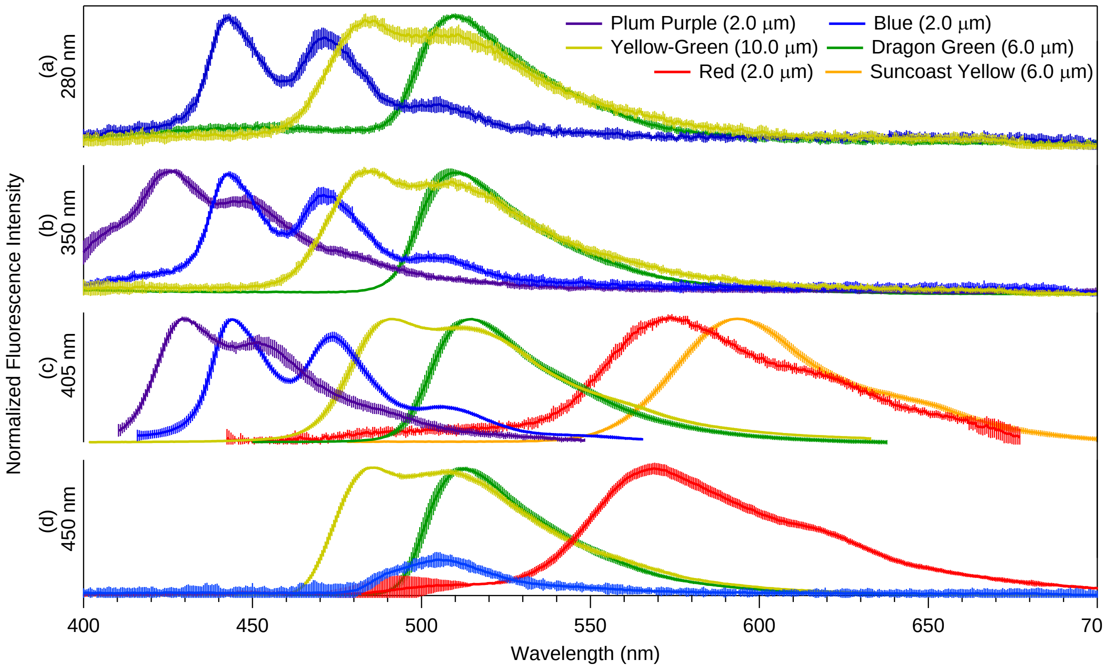


Figure 3.4. Fluorescence emission spectra of PSLs shown for four excitation wavelengths: 280 nm (a), 370 nm (b), 405 nm (c), and 450 nm (d). Spectra normalized to maximum peak intensity of unity. Each spectrum shown as an average of 17 particles, with vertical error bars representing the relative standard deviation of the intensity.

Figure 3.4 shows that the positions of spectral peaks are highly consistent across excitation wavelengths where a given fluorophore is active. The relatively small wavelength uncertainty of ca. ±2.5 nm for emission spectra collected from individual particles at a given excitation wavelength and ca. ±0-7% in relative standard deviation of intensity (thickness of traces) shows that highly reproducible spectral properties can be achieved by the simple technique. Figure 3.4 also shows reproducibility (i.e. ± 3.5 nm) of peak location across excitation wavelengths, showing that the fluorescent emission wavelength is essentially independent of excitation wavelength (i.e. Kasha's Rule).

It is important to note that even though absolute intensity is not represented in Figure 3.2, the technique is able to reproducibly detect relatively low levels of fluorescent emission. For example, the Red fluorophore is reported by the manufacturer to be fluorescent only at longer excitation wavelengths, marginally fluorescent at 450 nm, and completely non-fluorescent at λ_{Ex} < 450 nm (Technical Support Spectra Documents, Bangs Labs; Fishers IN). In contrast to manufacturer literature, Figure 3.2 shows reproducibly detectable spectra of the Red PSLs using both the 450 nm and 405 nm excitation sources. In some cases (i.e. Blue PSL under 450 nm excitation), only a partial emission spectrum is present, because emitted photons with wavelength below the cut-off are blocked by the optical filter.

3.2.4 Particle Sizing Methods

Ability to measure particle size is also an important factor in the identification and differentiation of biological particles, due to the characteristic nature of particle size within a biological species. Particle size measurements were done in two distinct phases: The 2018 phase and the 2019 phase. The sizing performed in the 2018 publication for Optics Express was a rudimentary method utilizing a single parameter and measurements from polystyrene latex spheres. Later measurements utilized a more complex ellipse measurement. Both are described in this section.

To measure the sizing properties of the instrument initially, six samples of Yellow-Green (YG) PSLs (Polysciences, Warrington, PA), ranging in size from 0.75 to 25 μm, were analyzed. Three different methods were utilized for particle sizing: (1) by directly measuring the number of pixels across a single axis of a given particle; (2) by measuring the height of the streak of dispersed light using white light excitation; and (3) by measuring the height of the streak of dispersed light using fluorescence dispersion from 405 nm excitation. The bounds of an individual particle or streak were defined as the point at which the observed light intensity dropped to 50% of its peak height (h50), in order to avoid sizing problems introduced by CCD saturation as particle size increased. The results of this analysis are summarized in Figure 3.5 and suggest that each sizing method can independently estimate the size of an individual particle in absolute units. For convenience, all sizing measurements discussed after this point were measured using the third method (fluorescence streak).

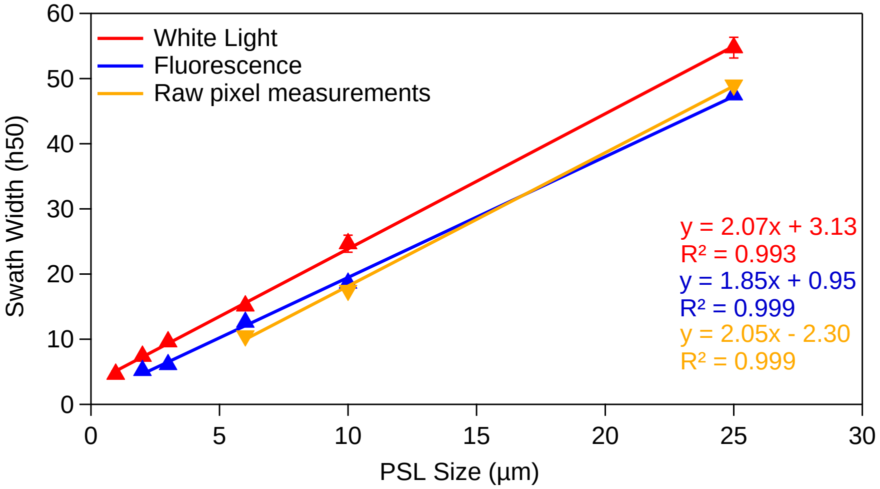


Figure 3.5. The h50 width of white light reflection (red), fluorescence dispersion width (blue), and raw, undispersed particle measurements (green). Markers represent average of 10 measurements with standard deviation shown as error bar. Where error bar is not visible, bar height is smaller than marker size. Lines show linear fit of all data, with line equations and R^2 displayed with corresponding color.

The particle sizing method as previously described utilized the width of the fluorescence swath measured by the sensor (Figs. 3.6a, b) as a proxy for particle diameter

(Swanson and Huffman 2018). The profile across the swath of light is extracted (red curve, Fig. 3.6b), and a Gaussian distribution is fit to the profile (black curve, Fig. 3.6b). The full width at half the maximum peak height (FWHM) is then taken as particle size. This method provides accurate sizing for spherical, homogeneous particles such as polystyrene spheres used for sizing calibration, as was shown previously (Swanson and Huffman 2018). The FWHM method can lead to sizing errors for particles with non-spherical morphology or inhomogeneous mixing of fluorophores, however. The swath of light diffraction through the grating is approximately the same height in the vertical dimension as the particle, but only with respect to the orientation of the particle on the stage. Oblong particles (i.e. aspect ratios higher than 1:1) can have any orientation, and so monodisperse particle of oblong shape exhibit a wide distribution of particle sizes. Additionally, particles that exhibit inhomogeneous composition or that contain areas with weak fluorescence (i.e. pollen grains with air pockets or lower fluorophore density) can show variations in fluorescence profile across the CCD image (e.g. bimodal distribution in Fig. 3.6b). Thus, the profile quality can vary also as a function of material composition, and calculated size will not be accurate.

To reduce sizing uncertainties from the effects mentioned above, an updated technique was developed to directly measure and record particle size and shape using the image of the particle. Raw calibration images are collected using simultaneous illumination by red and blue lasers, with the grating in place to disperse red and blue scattering points from one another. Particles are detected in calibration images automatically using Igor Pro analysis software (Wavemetrics; Lake Oswego, Oregon), searching the image at locations matching the diffraction angle from the red laser. A numerical threshold (T1), generally between 25- and 100-pixel intensity units, is applied to the images to convert light intensity values to binary. The T1 value is assessed for 1-2 particles per experiment by visually comparing with the original calibration image to ensure the size of the binary mask qualitatively matches the tested particle. The number of counted pixels within each particle ellipse are counted, and particles below a chosen threshold (T2) corresponding to approximately 10 µm in diameter are filtered out to limit detection of small particles and scattering artefacts. For each detected particle, the major and minor diameters are recorded using properties of the measured ellipses (Fig. 3.6c).

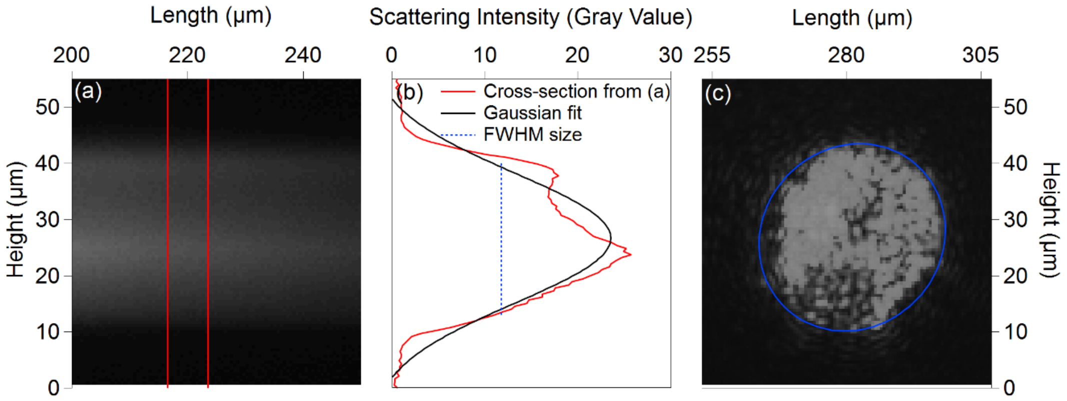


Figure 3.6. Particle sizing methods. (a) Example of fluorescence swath image used in previous method to measure particle size. Vertical red lines represent region in which light intensity was averaged in horizontal dimension. (b) Transect of light intensity from (a) and Gaussian fit. FWHM represents particle size measurement (26.5 µm). (c) Blue ellipse shows scattering image of an individual particle under red laser illumination. Particle size from (c): major and minor axes (38 and 36 µm) and Y and X dimensions (35 and 34 µm). Full images for (a) and (c) shown in Appendix Figure A3.

A new baseline subtraction is performed using a simple line curve subtraction utilizing Igor’s curve fitting functionality. Curves are fit prior to increases in spectral signal on either end (ex: 470 nm for the 450 nm excitation, corresponding with the optical filters, or 570 nm for the 280 nm excitation, corresponding with passing second order effects and increasing noise).

3.3 Factors Affecting Spectral Calibration

Normalized fluorescence spectra (e.g. Figure 3.4) can provide qualitative information about the presence of certain fluorophores. However, it is critical to be able to differentiate between particle types by accurately measuring absolute fluorescence intensity as a function of emitted wavelength and resolving small differences in intensity. Broadly, factors that influence the detected fluorescence intensity will be introduced below and can be organized into three general areas: (1) Effects of particle size, (2) Effects of camera settings, and (3) Effects of instrument optics. Several of these groups of factors contain a variety of individual effects that each scale with a different variable and that modulate the detected signal. Each of these factors will be isolated and investigated individually. Once measured, influence from each can be conceptually combined a single calibration that adjusts the intensity detected from each particle as a function of all available parameters. To test the effects of each set of individual parameters, the integrated fluorescence intensity of YG PSLs was measured by the instrument. These tests were performed with excitation wavelength blocking in place, but without inclusion of the grating. This allowed all fluorescent light reaching the detector from each particle to appear as a single dot representing each particle. For each experiment, Yellow-Green PSLs ranging in size from 1.5-4.0 µm were deposited onto a glass slide via impaction from the aerosol phase. This deposition method resulted in a consistent layer of PSL particles with sufficient particle number density within the viewing area. A single excitation source (405 nm) was used for each measurement in this section, and it is expected that these results will apply to each excitation source.

3.3.1 Effects of Particle Size

The intensity of fluorescence emission is strongly affected by particle size. For example, a particle with a larger number of fluorophore molecules or with higher fluorescent quantum yield may appear more intense than a smaller particle with fewer or weaker fluorophores. To be able to adequately characterize bioparticles that can naturally exhibit relatively wide distributions of properties within a given species, it is important to independently measure particle size and fluorescence intensity normalized per unit surface area of the particle. To measure this specific relationship within the instrument, five sizes of YG PSLs (1 – 25 µm) were illuminated by the excitation laser using a constant camera exposure time (0.87 seconds), camera gain (0.26, default value), and particle positioning within the viewing area. Figure 3.7a shows the average emission intensity as a function of PSL physical diameter. It is noteworthy that the observed intensity varies linearly with particle size. This observation is in contrast to the relationship that has been shown previously for real-time UV-LIF instruments for which the total intensity of measured fluorescence scales approximately with the 2nd to 3rd power of particle size (Hill et al. 2015; Sivaprakasam et al. 2011). In the collected image [e.g. Figure 3.7b], the X-dimension represents wavelength, the Y-dimension represents the width of the corresponding particle in the orientation observed (assuming no agglomeration), and the total intensity of fluorescent light is recorded for each pixel. The streaks of light are then averaged in the Y-dimension to produce fluorescence emission spectra as emission intensity vs wavelength. Since the spectra are averaged over an integration window in the Y-dimension, the cumulative influence of emission in this dimension is also averaged away. Particle width is directly correlated to both the X- and Y-dimensions appearing within a single image. This X-dimensionality also directly correlates to the perceived intensity of an individual spectral streak, leaving the raw intensity of the particle singly dimensional. It should be noted that particles may either saturate the fluorescence detector if large, brightly fluorescing particles are analyzed or may be below detection thresholds if small, weakly fluorescing particles are observed. These limitations can be mitigated by tuning the detector gain to optimize the collection of spectra from particles of interest.

3.3.2 Effects of Camera Detection Settings

Camera exposure time and gain are the two variables that can be controlled with the Lumenera camera software. Both variables impact the intensity of collected light, and thus need to be controlled for. To examine the relationship between intensity and each variable, 2.0 µm YG PSLs were observed. The 2.0 µm PSLs were utilized to avoid issues with CCD saturation. While holding the camera gain constant (0.26), the fluorescence intensity from a single PSL particle was measured as a function of varying exposure time [Figure 3.7b]. Separately, while holding the camera exposure time constant (6.28 seconds), intensity was measured as a function of varying gain [Figure 3.7c]. Both sets of measurements show a linear relationship over a range of values that do not promote detector saturation.

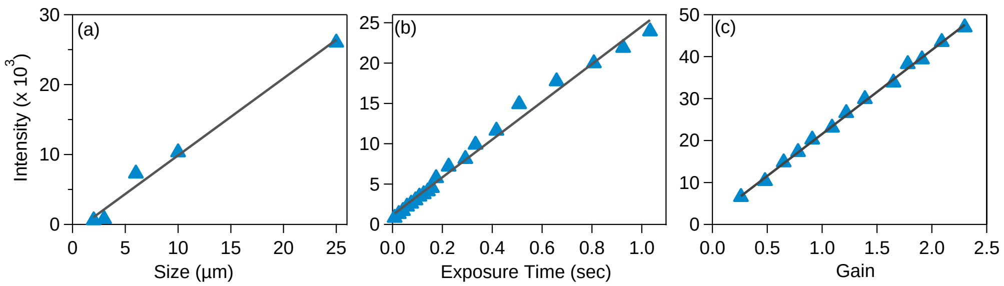


Figure 3.7. Comparison of parameters that influence measured emission intensity: particle size (a), camera exposure time (b), and camera gain (c). Lines show linear fits of measured values.

3.3.3 Effects of Instrumental Optics

After controlling for effects caused directly by the particles and within the camera software, a more complex set of variables collapse into a broad set of effects related to instrument optics, which cannot easily be separated or independently investigated. These effects can be organized into three groups: (i) the profile of the illumination beam, (ii) perturbations in collection optics (lenses, grating), and (iii) perturbations on the CCD surface.

Each individual excitation source presents a unique beam profile as a function of differences in overall source power, distance from the stage, angle of beam incidence, as well as beam shape and consistency. Differences in illumination power density influence the consistency of excitation energy impinging on a given particle. Laser beam profiles are generally much narrower than LED beams, which also have lenses placed in front of the source for focusing. Without correcting for illumination power-related differences in emission intensity, differences in emission spectra between two excitation sources are dominated by illumination strength rather than fluorescent properties of the particles themselves. To roughly estimate the illumination power density at the microscope stage, a fluorescent card (VRC1; Thorlabs) was placed in the plane of the microscope stage. The area of illumination was traced with a pencil, and a total surface area was approximated (to i.e. ± 5%). The total source output power was estimated (order of magnitude) for each source as the manufacturer-reported power as listed in Section 2.2. The power density was estimated for each source as the reported power divided by the calculated surface area: 0.004 mW cm^{-2} (280 nm), 0.5 mW cm^{-2} (350 nm), 4 mW cm^{-2} (405 nm), and 0.3 mW cm^{-2} (450 nm). These values are meant to be a first approximation of power density enabling a rough correction factor to allow comparison of emission intensities determined primarily by particle fluorescence properties. For example, the power density of the illumination beams is unlikely to be consistent as a function of location, with intensity generally decreasing radially outward, and imperfections in lenses will also manifest as beam perturbations.

Defects in optical components also affect the observed intensity of light collected at the camera and are expected to be more pronounced using inexpensive components, as in the present case. For example, it was observed that differences in particle position on the viewing area cause differences in focus due to a positive spherical aberration, and so it is impossible to maximize focus and resolution for all points on the emission spectrum (Smith 1922). The lenses also exhibit certain hot spot regions where additional light is focused. Several of these defocusing issues can be corrected with achromatic lenses, but these are more expensive than the components presently used and violate the goals of utilizing inexpensive optical components that could allow eventual wide scale production of an inexpensive device.

Standard camera CCDs are often practically limited to detection of photons with wavelength longer than 400 nm, as is the case for the camera used here. CCD detection can also be limited by dark noise. Cooling of the CCD chip could significantly decrease noise and increase both signal-to-noise ratio and sensitivity. A non-cooled CCD camera was chosen in this case to investigate whether spectral results would be reliable using the presently applied quality of components. Inexpensive CCDs can also have decreased detection efficiency across the spatial extent of the chip. For example, the camera employed here exhibits a clearly noticeable ring of weaker detection around the edges of the chip, with more extreme effects on the edges of the long-axis.

As mentioned, the three classes of optical effects manifest in a specific pattern of observed emission or scatter intensity as a function of placement on the particle stage. Instead of rigorously calibrating for each effect, a method was developed to empirically account for all three individual factors simultaneously. Images of a single batch of 2.0 μm YG PSLs were collected by illuminating particles with the 405 nm source and using a constant camera gain (0.26) and exposure time (0.0083 sec). The transmission grating was not utilized in this case in order to increase the density of particles observable by removing the dispersed spectral streaks. After acquiring a single image, the substrate was moved, and another image was acquired. Individual images were overlaid, and the collection process was repeated eight times until the sum total number of particles analyzed was 806. Particle position and average emission intensity of each individual particle was calculated, as shown in Figure 3.8a. Darker red colors toward the center of the viewing area indicate brighter emission intensity, and green colors on the edges show weaker measured emission. Because all the particles are identical in size and fluorescent properties the only differences in observed fluorescence emission are due to the combination of the three classes of instrumental optical effects discussed above. From a basic perspective, the particles on the left side and at the edges of the viewing area appear to emit more weakly as a combined function of weaker excitation power, perturbations from optical components, and limitations in detection efficiency at the edges of the CCD. To adjust for these combined effects, a map of observed emission intensity was calculated as a function of location in the viewing area (Figure 3.8b) using a Voronoi interpolation function applied to the composite image (Figure 3.8a). The Voronoi interpolation weights the space between a single point and its neighbors using a tessellated partitioning of the pixel plane and then interpolates the space between each individual point based on the weighted factors (Sibson 1981). The resulting surface represents the detectable emission intensity that would be expected for particles at each location in the viewable area.

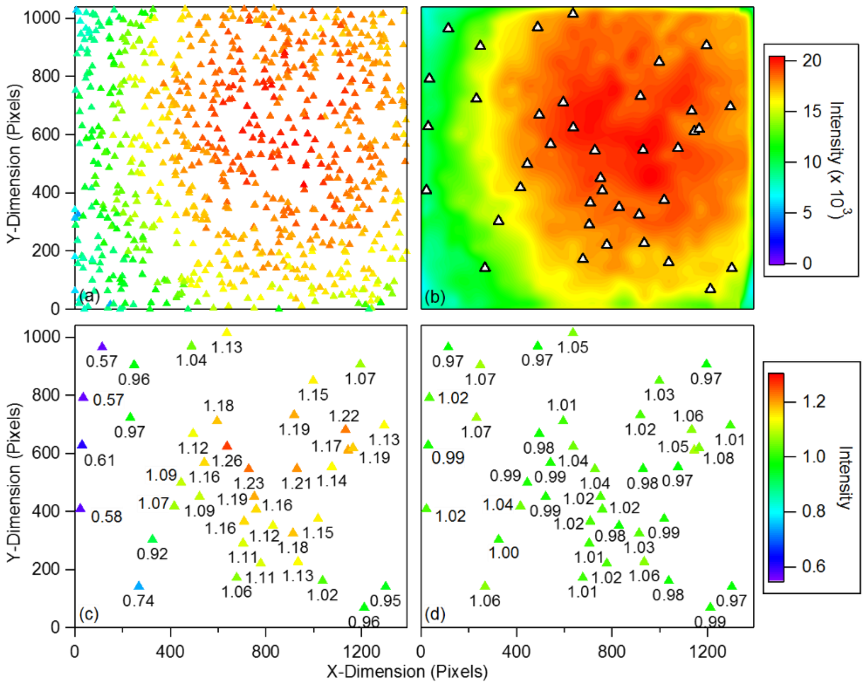


Figure 3.8. Calibration for intensity effects based on the spatial location of particles in the CCD viewing area (1392 x 1040 pixels). Eight non-dispersed images using 2.0 µm YG PSLs combined into a composite image of 806 particles (a). Image plot showing interpolation of values using composite image (b), including the location of test particles on the map (white triangles). The pre- and post-calibration values of the test set are shown in (c) and (d) respectively and are both normalized to the average intensity of surface map shown in in 5b. Normalized emission intensity of each test particle pre-calibration has an average of 1.052 ± 0.186, with a post-calibration average 1.011 ± 0.031 (d). Color scales represent raw intensity for (a) and (b), and normalized intensity for (c) and (d).

The quality of the intensity calibration was tested by interrogating a new image with 39 particles, which are shown overlaid onto the map with triangular markers (Figure 3.8b). The emission intensity values observed across the initial image presented a relative range of 19% (represented as relative standard deviation, RSD). To correct for optical effects discussed above, the observed intensity of each particle was divided by the interpolated values from the map, thus normalizing the emission of each particle to approximately unity. In this way, individual particles whose normalized intensity is 1.0

imply that the combined optical effects have been corrected for. After this normalization the average emission intensity was 1.011 with a RSD of 3.1%, representing a 6-fold improvement in the relative precision of the measurement, as shown in Figures 3.8c and 3.8d. Particles within 10 pixels of the edge (or 0.7- 1.0% of the surface area) were not included in this analysis, because of perturbations at the extreme edge of the CCD detector. By performing this normalization we established a method to significantly reduce the influence of optical effects introduced by inexpensive components.

3.3.4 Noise Filtering

Spectra utilized for clustering trials discussed in thesis chapters 4 and 5 were filtered to remove noise in the tails of the spectra, with intensity values less than approximately 0.1 arbitrary units. This noise filtration restricts the emission spectral range to 440 – 620 nm following 350 nm excitation, 440 – 650 nm following 405 nm excitation, and 450 – 670 nm following 450 nm excitation. Emission spectra following 280 nm excitation were not affected (range 400 – 560 nm).

3.4 Conclusion

This chapter demonstrates improvements to the instrumentation by using a more sensitive camera sensor and the introduction of more excitation sources to widen the breadth of accessible information. The camera introduced here is a monochrome camera that lacks additional optical filters and thus allows for more robust control of spectral collection via exposure time and gain settings. The addition of three new excitation sources quadruples the data obtained through the instrumentation, leading to fluorescence collections of bioparticles similar to a chopped-up excitation-emission matrix, what we call a “pseudo EEM.” These advancements allow for a multitide of imfornation to be obtained for an individual particle, allowing for comparisons within and between particle types.

Chapter 4: Clustering Strategies for the Classification of Pollen

The information present in this chapter was previously published in Swanson and Huffman, 2018 for section 4.4.2. All other clustering data was published in Swanson and Huffman, 2019, currently in review.

4.1 Introduction to Classification Strategies

Individual measurement observations from different sources will likely contain degrees of varying response signals depending on composition, size, or other properties inherent to what is being observed. Investigation of differing types of signals can be performed using various clustering and classification algorithms. Generally, a clustering algorithm takes input data for these observations, and attempts to group them based on relative similarity. This may be done a variety of methods, including connective-based (Murtagh and Contreras 2017), centroid-based (Taillard 2003), distribution-based (Xiaowei Xu et al. 1998), and density-based clustering techniques (Daszykowski and Walczak 2010). Each of these methods utilize different algorithmic framework, such as grouping observations by single-observation variable connectivity or by examining guassian distributions of observations, though are ultimately based on grouping similar observations into clusters. A simplified diagram of an extremely ideal scenario of this concept is shown in Figure 4.1. Assessing the efficacy of these techniques is also well studied, and similarly as complicated as the clustering itself, and can be done in two common ways: internal (based on information intrinsic to the data) or external (based on knowledge about the data prior to assessment) (Rendón et al. 2011).

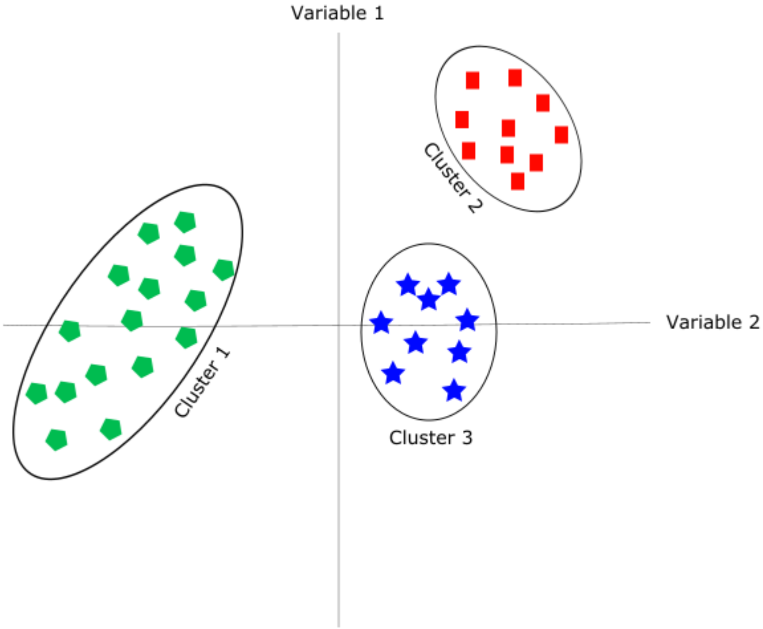


Figure 4.1. A simplified cartoon diagram of a two-variable clustering scenario for a set of observations. Each cluster is highlighted via a circle, with different colored symbols representing each type of observation. This represents an extremely ideal clustering scenario.

Clustering algorithms can also be separated into supervised or unsupervised techniques, where unsupervised refers to a technique that doesn't label data a priori. Supervised analysis involves the creation of a training data set to compare against new observations (Mohri et al. 2013). Unsupervised methods can be treated similar to supervised methods, since inputting known data will still provide an unbiased clustering result as if the data was unlabeled (i.e. observation identity wasn't known prior to clustering). Generally, these techniques can be extremely insightful when used in conjunction. Supervised algorithms, however, can present problems when not specifically examining trends in known data. Limits arise when predicting new types of observations to models that do not contain these types of observations. This is because these models absolutely will predict this new type to the model, though simple response outputs (e.g. a model made of “A, B, C,” observation types will still predict type “G,” without it being a part of the model). Therefore, care needs to be taken in making assumptions about the results from these models.

4.2 Classification Strategies Applied to Bioaerosol Analysis

A variety of multivariate analysis algorithms have been applied to the differentiation of spectral data from UV-LIF and other bioaerosol sensors (Huang et al. 2011; Pinnick et al. 2004). Algorithms can be divided into supervised or unsupervised classification techniques, where supervised techniques require prior input of data to train clusters, whereas this is not required for unsupervised techniques (Mohri et al. 2013). Unsupervised methods can thus be attractive to analyze particles from ambient observations, because no prior input is needed and so properties of test data do not bias results. For example, *k*-means clustering (unsupervised) was first applied to atmospheric aerosol data at least as early as 2004 (Erdmann et al. 2005), and has also been applied more recently, including with respect to particulate matter investigated using aerosol time-of-flight mass spectrometry and sun photometry (Elangasinghe et al. 2014; Rebotier and Prather 2007; Knobelspiesse et al. 2004). Unsupervised hierarchical agglomerative clustering (HAC) has also been frequently applied to study UV-LIF bioaerosol data, e.g. applied to WIBS data (Forde et al. 2019; Savage and Huffman 2018; Crawford et al. 2015; Robinson et al. 2013), and to fluorescence spectra from instrumentation that acquires LIF spectra at higher resolution than the WIBS (Könemann et al. 2019a; Zhu et al. 2015; Pan et al. 2003). Supervised clustering techniques can be effective analysis tools, especially when data is labeled (i.e. the identity of an observation is known). Ensemble methods, a subset of supervised methods that combine multiple learning algorithms in succession, can generally provide higher classification accuracy than unsupervised methods (Rokach 2010). The random forest (RF) classification technique is an ensemble algorithm that has previously been shown effective in differentiating bioaerosols (Ruske et al. 2017). The RF technique utilizes many parallel decision trees, each of which performs classifications via a series of decision nodes, by random bootstrap sampling of input variables. Tree decisions are then averaged to match an observation to the best matching input cluster. The RF technique has been shown to produce results with intermediate-quality separation accuracy (>74% for laboratory generated aerosols), but without requiring high computing power (Ruske et al. 2017). Gradient boosting (GB), another ensemble method, similarly creates small decision trees, with the exception that all variables are initially weighted equally and examined (Friedman 2011). The developed trees are then analyzed for variables that lead to misclassifications, and variables are re-weighted in order to circumvent misclassifications (Friedman 2011).

Within the last few years, the development of small and relatively inexpensive instrumentation for pollen detection has become more popular. In most of these cases, the physical principles of detection are based on light-scattering, pattern recognition, or holography, using advanced analysis computing to differentiate pollen types using field-portable instrumentation. One such prototype sensor generates diffraction holograms associated with individual particles, and deep learning techniques are then utilized to process and subsequently classify, or label, the measured particles from the hologram (Lee et al. 2011)**.** The sensor was shown to successfully separate a laboratory-generated mixture that included three species of pollen (Bermuda grass, oak, and ragweed), two fungal spore types, and common dust, with a classification accuracy of 94% (Wu et al. 2018). Another recently available commercial sensor is the Pollen Sense™ (Pollen Sense, Salt Lake City, Utah), which is a portable and relatively low cost (~$6,000) sensor that utilizes a combination of visual microscopy and image analysis techniques to identify pollen types as well as other large particles (http://pollensense.com).

Building upon a long history of pollen research using UV-LIF techniques and adding the goal of small sensor deployment, we previously developed an inexpensive mobile-platform-oriesensor with intended application toward pollen and fungal spore classification (Swanson and Huffman 2018; Huffman et al. 2016). Previously published work utilized only a small fraction of the acquired spectral data for analysis and so particle differentiation capabilities were limited. To more fully investigate the accuracy of the sensor with respect to pollen detection, in this study we first present improvements made to the sensor and to the image-processing procedure in order to utilize higher quality fluorescence spectra. Using this updated process, data was collected from 25-30 particles for each of eight different pollen species. Four clustering techniques (*k*-means, HAC, GB, and RF) were compared with respect to their ability to differentiate individual pollen grains from different species. RF and GB algorithms classified pollen with the highest accuracy with respect to the input data, and the algorithm parameters were further refined to optimize pollen separation. The clustering applications discussed here utilizing spectral data from this instrument show how optimizing the clustering process can improve particle differentiation with respect to the specific sensor and also have broad application to a growing number of techniques that utilize pollen data collected from other UV-LIF instruments.

4.3 Pollen Data Collection for Data Presented in Chapter Four

For Section 4.4, fluorescent emission spectra were acquired from four species of pollen grains: a) *Ambrosia trifida* (Giant Ragweed; weed; ~33 µm) b) *Alnus glutinosa* (Black Adler; tree; ~40 µm), c) *Zea mays* (Maize; grain; ~133 µm), d) and *Boussonetia papyrifera* (Paper Mulberry; tree; ~21 µm). All pollen was purchased in dried form from Bonapol (České Budějovice, Czech Republic). These four types of pollen were chosen to represent different classes of anemophilous pollen-producing plants and because it was expected that they would exhibit spectral differences, based on a previous study by Pöhlker et al.(Pöhlker et al. 2013). Additionally, Paper Mulberry and Giant Ragweed pollen were chosen due to their allergenic relevance. Maize and Black Alder were chosen to provide breadth in botanical taxonomy (i.e. grain and tree pollen, respectively).

Eight pollen species were chosen to represent a wide variation of plant species for investigation. Pollen were purchased from Allergon AB (Ängelholm, Sweden): Poa pratensis (Kentucky bluegrass; 011608102); from Sigma-Aldrich (Munich, Germany): Artemisia tridentata (big sagebrush; P9520); from Polysciences (Warrington, PA, USA): Broussonetia papyrifera, (paper mulberry; 07670); and from Bonapol (České Budějovice, Czech Republic): Ambrosia trifida (giant ragweed; 294-01-1-10), Betula pendula (silver birch; 134-04-1-13), Pinus strobus (eastern white pine; 225-02-1-15), Solidago Canadensis (Canadian goldenrod; 262-05-1-12) and Taraxacum officinale (common dandelion; 241-01-1-07). All pollen samples were deposited onto a pre-cleaned microscope slide by shaking a small amount of pollen out of a plastic bag or by impacting using an aerosol collector and pump.

4.4 Application of Truncated Data; Supervised *k*-means

4.4.1 Truncated Clustering Method

To aid in the quantitative differentiation between particle types, the *k*-means clustering algorithm was utilized (MacQueen 1967). For this clustering trial, the full range of emission spectra was not utilized in order to simplify the process as a proof-of-concept. Particles were represented individually in the algorithm as nine input parameters: particle size as well as the wavelength and intensity of emission spectral maxima from each of the four excitation sources. *k*-means clustering was performed on the open source software RStudio (RStudio Team 2016), utilizing an internally available statistical package. Data values were scaled prior to clustering using the automatic function within RStudio to weight each factor equally. The *k*-means algorithm used these nine values to develop a cluster representative of each pollen species.

4.4.2 Truncated Clustering Results

Ten individual particles were analyzed for each species at each of the four excitation wavelengths. For this exercise, camera gain (0.26) was held constant, and exposure times for each excitation were conserved across all images for a single species to ensure visible, but unsaturated, spectral streaks. Spectra were acquired individually by placing the particle at the center of the viewable area, and each individual emission spectrum was normalized by dividing by the average illumination power density mapped at the location of the particle and by the measured values for exposure time and individual particle size. Figure 4.2 shows the results of these 160 spectra. There are clear differences between individual particles within a species (individual panel), as expected between individual biological entities and due to differences in viability state, aging, or growth conditions. The differences between species types, however, are much greater than intra-species variability.

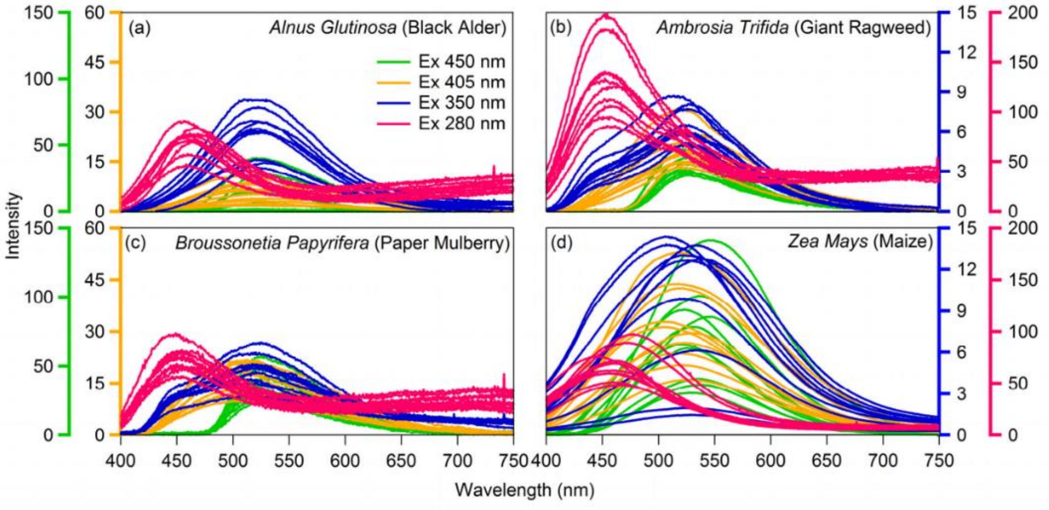


Figure 4.2. Emission spectra collected from four species of pollen: (a) Alnus glutinosa, (b) Ambrosia trifida, (c) Broussonetia papyrifera, and (d) Zea mays. Emission spectra colored by excitation wavelength and scaled relative to axis with matching color. Spectra are normalized according to details discussed in text. Ten particles of each species are shown. Spectral baselines increase at long wavelength for 280 nm excitation due to light noise caused by experimental set-up.

To quantitatively test the quality of separation between pollen species types, results from all 40 individual particles were analyzed using a supervised *k*-means clustering algorithm. As a first attempt, emission spectral data were summarized as the peak location (wavelength) and intensity for each of the four excitation wavelengths. These eight values, as well as particle size, for each of the forty particles were input into the clustering algorithm. Despite the simplified inputs from single particle data, the algorithm separated the forty particles into four clusters, organized with 100% correctness. To illustrate the efficacy of the clustering itself, Figure 4.3 shows a three-dimensional section of the cluster data, arbitrarily representing three of the nine output dimensions (parameters). Individual particles are colored by cluster (or species type) and black markers represent averages of each member of a given cluster. By this representation, using only three of the nine dimensions the clear separation between pollen types and cluster centers is visible.

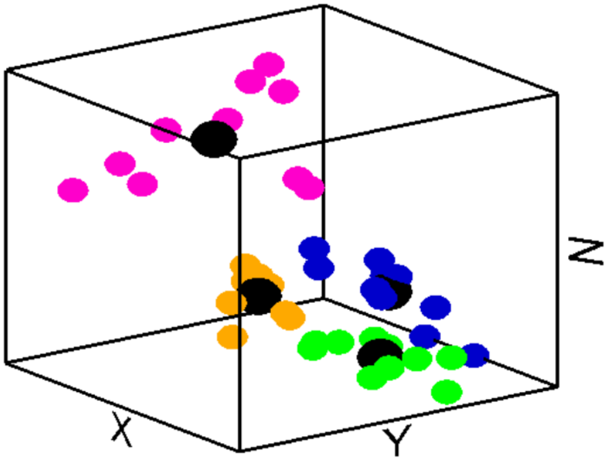


Figure 4.3. Four-cluster solution produced using *K*-means algorithm and represented in three dimensions showing separation ability of clustering process. Axes represent: wavelength (X-axis) and normalized intensity (Y-axis) of maximum emission peak following 405 nm excitation, and particle size (Z-axis). Colored dots represent individual particles from: cluster 1, Alnus Glutinosa (blue); cluster 2, Ambrosia Trifida, (green); cluster 3, Broussonetia Papyrifera (orange); and cluster 4, Zea Mays (pink). Black dots show center of each cluster.

These initial clustering results show the ability of the technique to reproducibly differentiate between classes of pollen despite limitations e.g. inability to detect emission spectra at λ_{Em} < 400 nm. In the future, clustering will be performed using full emission spectra collected for each particle. This will significantly increase the dimensionality of data used for the clustering algorithm and will be powerful in scenarios in which particle emission spectra exhibit multiple peaks or other distinct features. It is anticipated that clustering will be at least as able to differentiate between particle types, and we anticipate that species even more closely related will be differentiable. The results presented here are intended as a proof-of-concept for differentiation of particle types.

4.5. Pollen Classification Strategy Comparisons

There are many these tools available for clustering data, though four were chosen for this study: *k*-means, HAC, random forest, and gradient boosting. Two of these methods, *k*-means and HAC, are operated as unsupervised clustering algorithms. Unsupervised algorithms are beneficial in bioaerosol classification because ambient data collections will not be labeled (Robinson et al. 2013). However, unsupervised methods may be significantly harder to interpret for a similar reason, as well as due to the large variability in atmospheric particles (Hernandez et al. 2016; Crawford et al. 2015; Pinnick et al. 2013). Many supervised clustering techniques have various tools to predict new, unlabeled data on previously developed training sets (Rokach 2010). Both random forests and gradient boosting have this implemented, and have shown promising results in bioaerosol categorization (Ruske et al. 2017). This study contains labeled pollen data, and as such examines the efficacy of these various clustering tools against data obtained through the sensor. An individual particle is represented in each algorithm as a series of 1063 variables: The major and minor size axes, the aspect ratio, and the emission curves for each excitation from 400-700 nm, except for the 280 excitation which is cut off at 560 nm. For these three trail comparisons, the particle emission spectra were scaled by the z-score method, though the other three parameters were not, as the size variables are grossly outnumbered by emission variables at a ratio of ~353:1.

4.5.1 *k*-means clustering

The *k*-means clustering algorithm utilizes an iterative process of randomly choosing data observations (*k*) as cluster centroids (Hartigan and Wong 2006). The observations are then partitioned based on the cluster centroid into a Voronoi partition, and new centroids are calculated based on these groupings. The algorithm continues this process until it converges to a local optimum and the centroid values no longer change. The *k*-means clustering algorithm was previously explored briefly using data from the sensor (Swanson and Huffman 2018), using 4 pollen species and reduced input data (height and position of emission spectral peak maximum). This process showed the ability to separate broadly between pollen species with wide taxonomic differences as a proof-of-concept, however the use of simplified data does not facilitate differentiation of pollen based on subtle features in the emission spectra.

Cluster analysis using the *k*-means algorithm was performed here as a semi-unsupervised process, in contrast to the method applied previously (Swanson and Huffman 2018), where the technique was applied in an unsupervised manner. This means the cluster centroids were not pre-defined here, and only cluster number (k=8) was prescribed to the algorithm. The *k*-means algorithm was also previously shown to produce relatively high misclassification of ambient bioparticles detected by a WIBS due to the limited nature of the unsupervised method used within this algorithm (Ruske et al. 2017), but exhibited relatively high accuracy using data obtained from the sensor discussed here (Swanson and Huffman 2018). When allowed to iterate until optimal clusters are created, the *k*-means algorithm produces clusters with similar group size (Percy and Everitt 2006). This does not present problems for the data presented here, but may introduce errors when unknown numbers of pollen species are involved, e.g. in ambient samples (Geburek et al. 2012). The *k*-means clustering algorithm is available as a built-in statistical package for R (RStudio, Inc, Boston, MA).

4.5.2 Hierarchical Agglomerative Clustering

The hierarchical agglomerative clustering algorithm initially uses the number of clusters matched to the number of measured particles, and groups data by Euclidian similarity until a single cluster remains. The user is then required to choose the appropriate number of clusters based either on *a priori* knowledge of the number of particle types or by using HAC-specific tools such as the Calinski-Harabasz Index that examines inter- and intra-class distance ratios (Liu et al. 2010). Allowing the algorithm to determine the optimal cluster number can be powerful, because previously unknown properties can be revealed (Robinson et al. 2013). HAC analysis has been applied to single-particle LIF data with relative success (Pan et al. 2012; Pinnick et al. 2004). Labeled and unlabeled data can both be examined by these unsupervised techniques by removing data labels to treat all data as unknown. The HAC output can be visualized using a dendrogram, which shows distances between observations in a representative tree diagram and which will report particle grouping from the top down. The dendrogram can be chopped at the desired number of clusters (e.g. n=8) for the final classification solution. Several linkage methods exist for the HAC algorithm, including single, average, weighted, complete, and Ward's (Crawford et al. 2015). The ward.D2 method was used in this study, similar to a previous study in which pre-labeled data was clustered utilizing HAC (Savage and Huffman 2018). HAC is available in the fastcluster package, an open source tool for R.

4.5.3 Random Forest Algorithm

Random forest classification is a supervised ensemble algorithm that utilizes decision trees to group observations based on bootstrap sampling of the data (Breiman 2001). Decision trees classify observations by making individual node decisions to separate observations but can suffer overfitting by developing a model that memorizes the data. In some cases an individual decision tree may produce accurate results for the training data, but inaccurate results for the subsequent data being tested (Dietterich 1995). RF classification algorithms use many conditional decision trees that are developed in parallel, utilizing random variations of the variable inputs (Hothorn et al. 2004; Breiman 2001). The RF method allows for development of both over-fitted and representative trees. Using a large number of trees for analysis allows many of the trees to be developed simultaneously, the majority of which should represent the data accurately. Changes in decision tree population (forest) can affect classifications and utilizing the optimal number of variables compared at each decision node implies inherent trade-offs between developing trees that examine the variables properly and that memorize the data.

Random forests have been employed for bioaerosol analysis and showed similar performance to other supervised techniques such as GB and neural networks, but with lower computational burden (Ruske et al. 2017). Random forests have also been used e.g. for genetic mapping, which requires use of a large number of variables from the sample data (Bureau et al. 2003). RF classification was performed here using the open-source ‘party’ package within R. The cforest tool (conditional RF algorithm available in the ‘party’ package) uses unbiased processes in the decision making. Unlike the base implementation of random forest within R, cforest trees are initially developed, then conditional inference trees are fitted to the originally developed bootstrap trees, and the averaged observation weights from the trees are reported rather than simple average values from the bootstrap trees (Hothorn et al. 2015). These two differences result in predictive models that are more accurate, but more computationally expensive. For initial testing, the number of trees used here was held at 500, and the number of variables was left at the package default of five.

An individual tree for the RF model can be plotted in order to visualize the decision-making process within the algorithm. This was shown for the entire data set in Figure 4.4 as an example, in which the 243rd tree is represented from a 500-tree forest. This tree classified the 204 particles through separations at nine distinct decision nodes. For the first decision node, the tree displays the most important variable of 50 chosen randomly from 1063 input variables: emission intensity at 473 nm following excitation at 280 nm.

Particles with emission intensity >18.8 at this wavelength were classified into cluster 8, ascribed as *T. officinale*. This delineation was effective, because no other species contained particles with emission intensity > 18.8. For example, see relative differences in pollen species with respect to emission spectra following excitation at 280 nm (Fig. 4.5). Each subsequent branch of the tree shown in Figure 4.4 separates observations based on other variables until final clusters are formed.

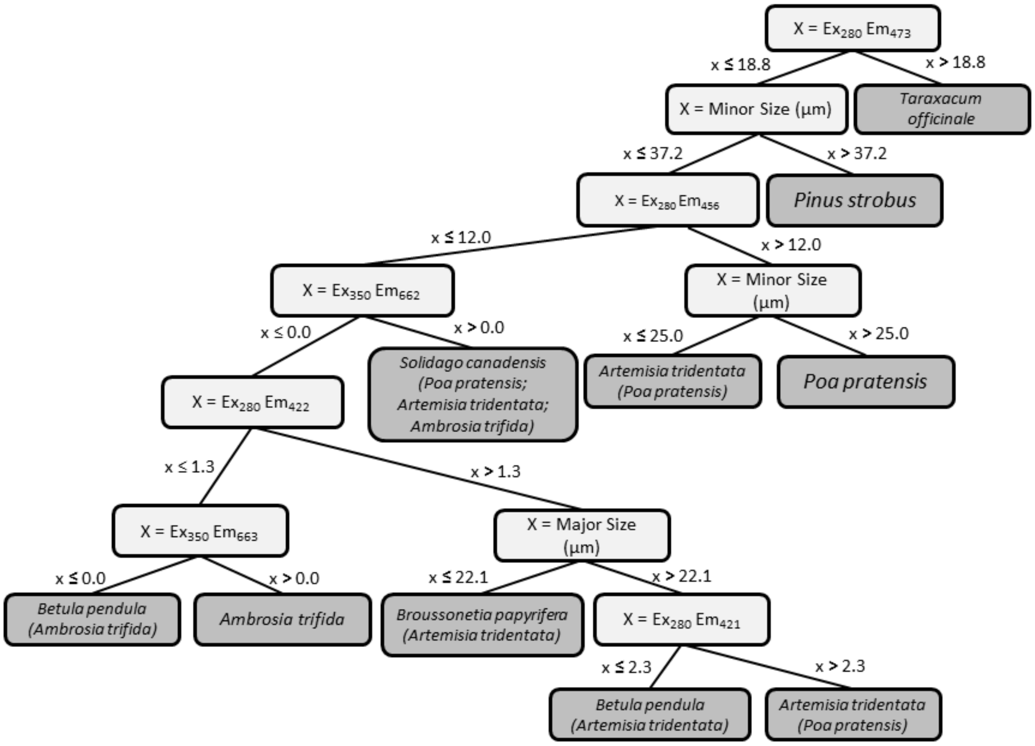


Figure 4.4. Single conditional tree (#243) from a 500-tree RF classification of entire pollen data set. Tree shows the decisions involved in classification of all 8 species. Highest impact variable listed in light grey, output nodes in dark grey. Main pollen species node stated in output node, with misclassified species listed in parenthesis.

The RF algorithm, as initially operated, provided higher classification accuracy than the two unsupervised algorithms, however improvements can be made by manipulating input parameters that affect development of the model. Increasing in the number of trees from 1 to 2000 provides higher accuracy, but the relative improvement diminishes as the model converges to optimal accuracy at ~500 trees (Appendix Fig. A4). Variable number examined per node can also be changed from a default value ('mtry' = 5). Increasing variables examined per node can ultimately allow development of identical (over-trained) trees, thus limiting the advantage a RF has over other techniques. To avoid this issue, mtry was left at 5. A 500-tree forest was used for the initial testing, and a forest of 1000 trees was used for Sections 4.5 and following.

4.5.4 Gradient Boosting Classification

Gradient boosting is a supervised ensemble classifier algorithm that uses smaller, weaker decision trees than the RF technique (Friedman 2001). The term "weaker trees" implies that they are developed with a single decision node to separate a fraction of the data per each weak tree. These weaker trees are used in an iterative fashion, as opposed to being developed simultaneously as in RF, where the overall model is re-trained to reduce mean squared error over the series of weak decision trees. Instead of randomly selecting variables, all variables are weighted equally and each iteration re-weights variables based on an exponential loss function. Variables are re-weighted based on misclassification performance in each decision tree, and the algorithm iterates repeatedly over a given number of trees. The algorithm process allows for a number of sequential decision trees to be made into a model that can accurately separate sections of data for each individual decision tree.

GB algorithms have shown relatively high accuracy with respect to sorting bioaerosol classes, though at higher computational cost than the RF algorithm (Ruske et al. 2017). Overfitting the data can occur frequently, so cross-validation can be performed automatically within the model through data sub-sampling (Friedman 2011). Sub-sampling allows the data to be split into *k* number of groups, which are then used to take *k-1* groups to develop a training mode used to test on the remaining group (James et al. 2013). The test-set error from sub-sampling is used to determine optimal tree iteration, which is the ideal position in the model to predict new data. GB was performed using a multinomial distribution; cross-validation folds of 10, and 500 trees, and is available from the 'gbm' package for R.

When improving the classification accuracy for GB, the risk of overfitting is present, though this can be mitigated with tools from the gbm package. Cross-validation can be used to develop an exponential loss curve that analyzes the difference in error associated with the training and testing sets (Appendix Fig. A5). Other gbm parameters were investigated, including shrinkage, interaction.depth, and n.minobsinnode. The effects each of these played on the data set were minimal, and default parameters led to accurate predictions. A model with 10-fold cross-validation was used in all circumstances, and the ideal iteration was used to predict data for Section 4.5.

4.5.5 Particle Misclassifications and Total Error

Classification results are shown in confusion matrices (e.g. Table 4.1), which visually describe the accuracy of classifications with respect to input category and output cluster. Particle misclassifications are described in terms of precision (false-positive) and recall (false-negative). Precision describes the ratio of particles incorrectly classified to a cluster (vertical misclassifications), to the number of particles correctly classified to the cluster. Recall describes the ratio of particles incorrectly classified from a cluster (horizontal misclassifications), to the number of correctly classified particles. A value of 0.0 for precision or recall variables describes misclassification, whereas a value of 1.0 describes correct classification. The precision and recall variables are used to calculate the mean *F* value for a species, e.g. averaged over a calculated cluster using the following equation (Buckland and Gey 1994):

$$F = \frac{(2\ x\ False\ Positive\ x\ False\ Negative)}{(False\ Positive + False\ Negative)}.$$

The *F* value thus allows representation of the misclassification vector of a cluster as a single variable and relates cluster accuracy. Results for the ensemble algorithms (RF and GB) were cross-validated using with a four-fold validation method, meaning 75% of the observations were utilized to develop a training set, and the remaining 25% were utilized to test it. This was to ensure there was no overfitting, as well as to test the accuracy of each classification model. Data was also selected in randomized order to add additional complexity to this validation.

4.5.6 Variable Importance

Ensemble algorithms offer two specific sub-routines that were used to analyze spectral data. The ‘predict’ (cforest random forest) and ‘gbm.predict’ (gbm gradient boosting) functions allow for both testing of training data, as well as predicting where new observations will be assigned. The new predictions can provide responses (particle assignment) or probabilities (percentage of similarity of an observation to any cluster) for an individual observation. The ‘variable importance’ function utilizes information from decision trees present in the algorithm to report the variables integral in correct classifications. Importance for a variable is reported as mean decreased Gini[1] (MDG), describing how the available data would be further misclassified by removing that single variable (Han et al. 2017; Strobl et al. 2007). MDG values and size variables were examined individually for each curve.

4.5.7 Reduction of Number of Optical Sources

Computational experiments were performed in which combinations of input variables (e.g. emission spectra associated with individual excitation sources) were removed in order to examine their relative importance for pollen differentiation. This test is analogous to physically operating the sensor without certain optical sources and helps indicate which sources are the least important to overall pollen classification and thus candidates for physical removal from the instrument. Sixteen individual trials were developed, all tested on an identical randomized data set. Reduction in data collection represents a tradeoff between increased observational collection and lowering the overall

[1] Term introduced by statistician Corrado Gini.

cost and time requirements in the analysis. These trials involved a cross-validation set similar to Section 4.5.3, which used 75% of the data to develop the training model and 25% of the data to test the model accuracy.

4.6 Classification Results and Discussion

4.6.1 Size and Spectral Characteristics of Pollen

Particle size and spectral information from 20 – 31 individual particles were collected for each of the 8 pollen species studied. By analyzing data averaged for individual species, patterns appear that aid discrimination and grouping (Fig. 4.5 and Appendix Fig. A7). Emission spectra from the 280, 405, and 450 nm excitation sources each exhibit a single, broad peak with a tail sloping to longer wavelengths, corresponding to fluorophore modes I, V, and VI, respectively. Emission spectra following excitation at 405 nm are weaker than for those following excitation at 450 nm for all species except *T. officinale* (Fig. 3H). This is explained by the fact that the 405 nm excitation crossed at a minimum between fluorophore excitation spectra peaking at ~350 nm and ~450 nm (Appendix Fig. A6). As a result, emission spectra following 405 nm excitation are dominated by the tail of the emission peak at 450 nm (III) rather than the tail of the emission peak at 520 nm (V). For spectra from these three sources, differences are apparent between species primarily due to the height rather than relative shape of emission peaks. Emission spectra from the 405 nm source can be broadly grouped according to peaks with high intensity (*T. officinale*), medium intensity (*A. tridentata*, *P. pratensis*, and *S. canadensis*), and low intensity (four remaining species). Emission spectra from the 350 nm excitation source, in contrast, show a broad peak at 460 - 540 nm representing two unresolved peaks. The first peak at ~470 nm (II) corresponds e.g. to phenolic compounds and the second at ~520 nm (IV) corresponds e.g. to carotenoids (Pöhlker et al. 2013). Previous studies have shown the relative intensity of mode II to be higher than mode IV for most pollen species. Spectra shown here exhibit lower intensity values for mode II, however, likely influenced by the optical filter used (435 nm long-pass filter; GG-435; Edmund Optics; Barrington, NJ) to filter the spectrally broad output from the 350 nm LED. The filter removes approximately 15% of light at 450 nm[2], and so the relative peak height of mode II is reduced and the shape of spectra following 350 nm excitation are qualitatively altered. Emission spectra following excitation by the 280 nm source shows the largest variations in peak height between species, spanning mean values between 3.8 and 30.6 (arbitrary intensity units), probing mode I related primarily to phenolic compounds. EEMs of pollen and collected from bulk biofluorophores suggest that the 280 nm source should promote fluorescence from proteins and aromatic amino acids (Pöhlker et al. 2012, 2013), peaking approximately at emission wavelength 350 nm. This mode is not visible with the present set-up of the instrument, because the efficiency of the silicon CCD used here for detection drops to near zero as wavelength drops below ~400 nm. Emission spectra from the 450 nm source exhibit variability, but within a narrower range than for other sources, and the location of peak maxima are nearly identical across all species.

Mean pollen size varied from 20 to 50 μm. Fluorescence intensity emitted from individual particles has long been shown to increase strongly as a function of particle size

[2] https://www.edmundoptics.com/document/download/352852

(e.g. Savage et al. 2017; Sivaprakasam et al. 2011; Hill et al. 2001). Differences in composition between pollen species also play important roles in observed fluorescence intensity, however. For example, *P. strobus* (Fig. 4.4e) exhibited large mean particle size (52 µm), but weak fluorescence intensity for 280, 405, and 450 nm excitation sources. The mean aspect ratio values of pollen species varied from 1.0 (*A. trifida*) to 1.6 (*A. tridentata),* with most species presenting mean values between 1.1 and 1.3.

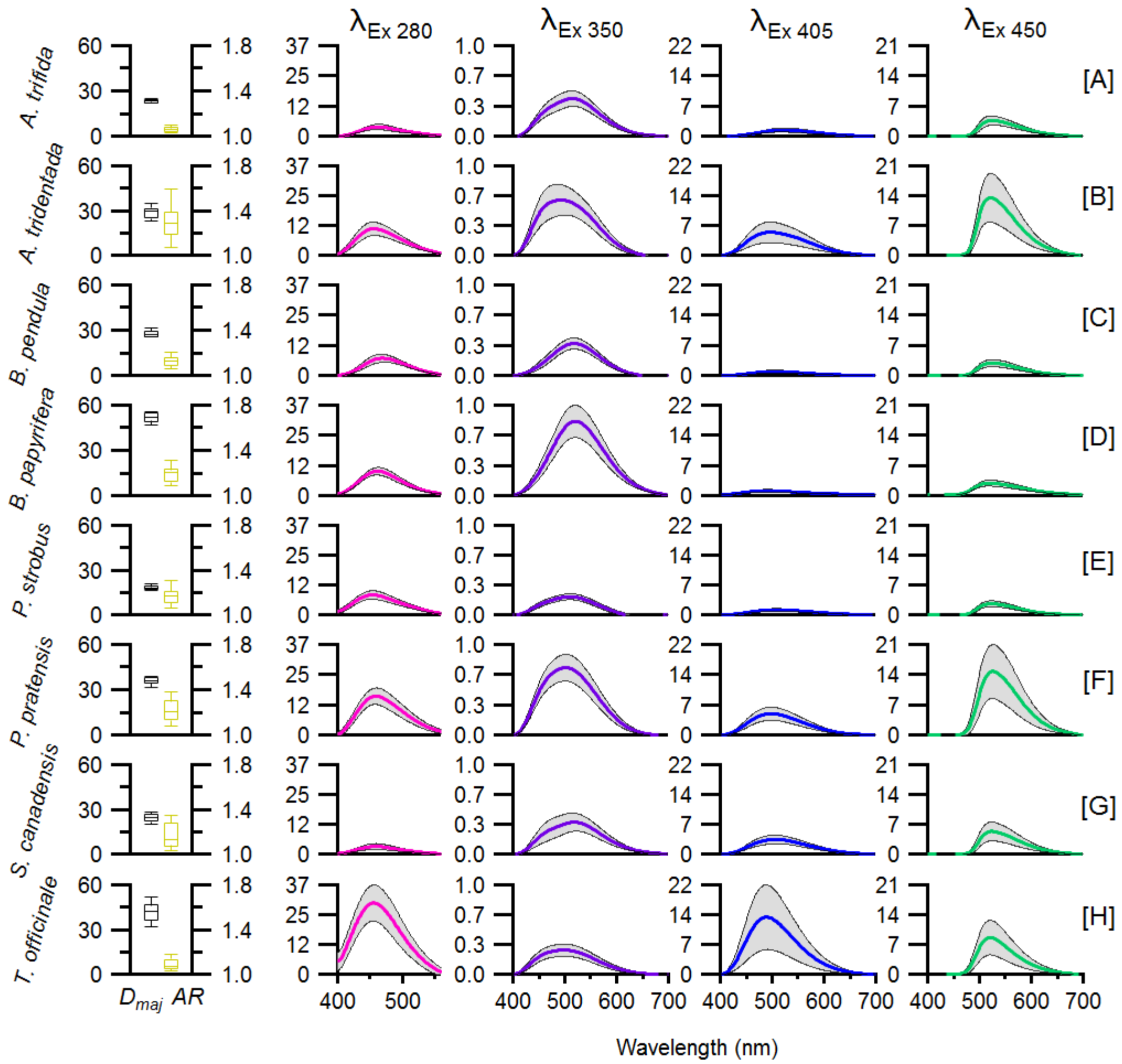
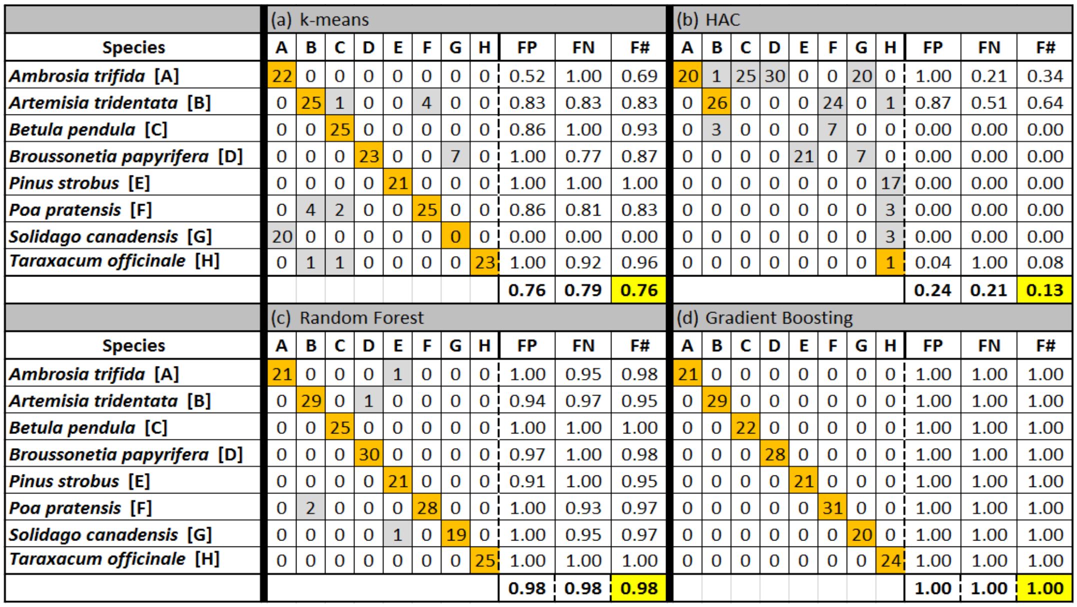


Figure 4.5. Measured properties of all pollen species analyzed. Major particle size axis (D_{maj}; black) and aspect ratio (AR; yellow) shown where box limits represent 25th and 75th percentiles, whiskers represent 10th and 90th percentiles, and center line represents median value. Remaining columns show emission curves following excitation at 280, 350, 405, and 450 nm. Center line of each spectrum represents mean value, grey region represents standard deviation of measurements. Emission intensity is conserved within a given column to aid comparison.

4.6.2 Comparison of Clustering Techniques

Average classification accuracy (*F*) for the four algorithms studied ranged from 0.13 to 1.00. The two supervised techniques (RF; *F* 0.96 and GB; *F* 1.00) significantly outperformed the two unsupervised techniques (*k*-means; *F* 0.78 and HAC; *F* 0.13), as summarized in Table 4.1. The GB algorithm classified the data to an average *F* of 1.00 but can also over-fit the data by developing trees that perfectly fit the training data. As a result, the *F* value can overestimate the true accuracy of the GB model. The RF algorithm correctly labeled particles with *F* of 0.98, corresponding to 2% error or 5 particles misclassified out of 204. The RF algorithm is not susceptible to over-fitting with default parameters, however, and so the *F* value more reliably represents assignment accuracy. In this case one particle from each of three pollen species was misclassified, and two additional *P. pratensis* particles were misclassified to the *A. tridentata* cluster. The spectra from these two species are relatively similar (Figures 4.4a and 4.4f), and so misclassification here is reasonable.

The average accuracy of the *k*-means algorithm was mediocre, with an *F* value of 0.76, and the HAC algorithm showed very poor accuracy with an *F* value of 0.13. This suggests that Euclidian distance between data points (HAC) may not sufficiently separate data in this case. The unsupervised methods consider all variables simultaneously by combining variables, in contrast to supervised methods that randomly sample subsets of the input data. The unsupervised process may result in variables that carry added weight if overlapped between species. Observations with large differences (e.g. a weakly fluorescent particle from one species and a highly fluorescent particle from another) may skew initially developed cluster centroids, making further groupings less accurate by increasing misclassification. Given that the RF and GB algorithms (even as operated without comparing observations to a training set) significantly out-performed the unsupervised algorithms, and because they can be further tuned to improve classification, only these two algorithms were utilized for further investigation here.



	(a) k-means											(b) HAC										
Species	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**FP**	**FN**	**F#**	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**FP**	**FN**	**F#**
***Ambrosia trifida* [A]**	22	0	0	0	0	0	0	0	0.52	1.00	0.69	20	1	25	30	0	0	20	0	1.00	0.21	0.34
***Artemisia tridentata* [B]**	0	25	1	0	0	4	0	0	0.83	0.83	0.83	0	26	0	0	0	24	0	1	0.87	0.51	0.64
***Betula pendula* [C]**	0	0	25	0	0	0	0	0	0.86	1.00	0.93	0	3	0	0	0	7	0	0	0.00	0.00	0.00
***Broussonetia papyrifera* [D]**	0	0	0	23	0	0	7	0	1.00	0.77	0.87	0	0	0	0	21	0	7	0	0.00	0.00	0.00
***Pinus strobus* [E]**	0	0	0	0	21	0	0	0	1.00	1.00	1.00	0	0	0	0	0	0	0	17	0.00	0.00	0.00
***Poa pratensis* [F]**	0	4	2	0	0	25	0	0	0.86	0.81	0.83	0	0	0	0	0	0	0	3	0.00	0.00	0.00
***Solidago canadensis* [G]**	20	0	0	0	0	0	0	0	0.00	0.00	0.00	0	0	0	0	0	0	0	3	0.00	0.00	0.00
***Taraxacum officinale* [H]**	0	1	1	0	0	0	0	23	1.00	0.92	0.96	0	0	0	0	0	0	0	1	0.04	1.00	0.08
									0.76	**0.79**	**0.76**									**0.24**	**0.21**	**0.13**
	(c) Random Forest											(d) Gradient Boosting										
Species	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**FP**	**FN**	**F#**	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**FP**	**FN**	**F#**
***Ambrosia trifida* [A]**	21	0	0	0	1	0	0	0	1.00	0.95	0.98	21	0	0	0	0	0	0	0	1.00	1.00	1.00
***Artemisia tridentata* [B]**	0	29	0	1	0	0	0	0	0.94	0.97	0.95	0	29	0	0	0	0	0	0	1.00	1.00	1.00
***Betula pendula* [C]**	0	0	25	0	0	0	0	0	1.00	1.00	1.00	0	0	22	0	0	0	0	0	1.00	1.00	1.00
***Broussonetia papyrifera* [D]**	0	0	0	30	0	0	0	0	0.97	1.00	0.98	0	0	0	28	0	0	0	0	1.00	1.00	1.00
***Pinus strobus* [E]**	0	0	0	0	21	0	0	0	0.91	1.00	0.95	0	0	0	0	21	0	0	0	1.00	1.00	1.00
***Poa pratensis* [F]**	0	2	0	0	0	28	0	0	1.00	0.93	0.97	0	0	0	0	0	31	0	0	1.00	1.00	1.00
***Solidago canadensis* [G]**	0	0	0	0	1	0	19	0	1.00	0.95	0.97	0	0	0	0	0	0	20	0	1.00	1.00	1.00
***Taraxacum officinale* [H]**	0	0	0	0	0	0	0	25	1.00	1.00	1.00	0	0	0	0	0	0	0	24	1.00	1.00	1.00
									0.98	**0.98**	**0.98**									**1.00**	**1.00**	**1.00**

Table 4.1. Clustering accuracy (*F*) comparison utilizing (a) *k*-means, (b) hierarchical agglomerative clustering, (c) random forest, and (d) gradient boosting classification algorithms using the entire pollen data set (204 particles). On the left of the dotted lines is a confusion matrix, in which correctly classified particles are highlighted in orange and misclassified particles in grey. On the right side, FP (false positive) represents the number of particles misclassified for that cluster in the vertical dimension, FN (false negative) represents the number misclassified in the horizontal dimension, and *F* represents the harmonic mean of these misclassifications for the cluster.

4.6.3 Detailed Comparison of RF and GB

Though the RF and GB algorithms performed well, the developed models may be over-trained. Prediction of new, labeled data (e.g. subsets of the data) to the model can thus be important to assess model performance. For both RF and GB, a series of five cross-validation trials were performed with the data sets, where 75% of the randomized data was used to create a training set and the remaining 25% was used as a test set. For most of the trials, RF and GB algorithms performed with similar overall accuracy. Averaged over the five trials, *F* was 94.8 ± 4.6 for GB and 93.6 ± 3.3 for RF (Fig. 4.6). GB shows higher *F* than RF during training, but the mean results are similar following testing. These results imply that GB can over-fit the training data despite built-in cross-validation and that RF training sets are more representative classification scenarios. Given the similar results between RF and GB, the similarities between training and testing accuracies exhibited by RF, and the lower computational expense (17x faster), further investigation was limited to the RF algorithm.

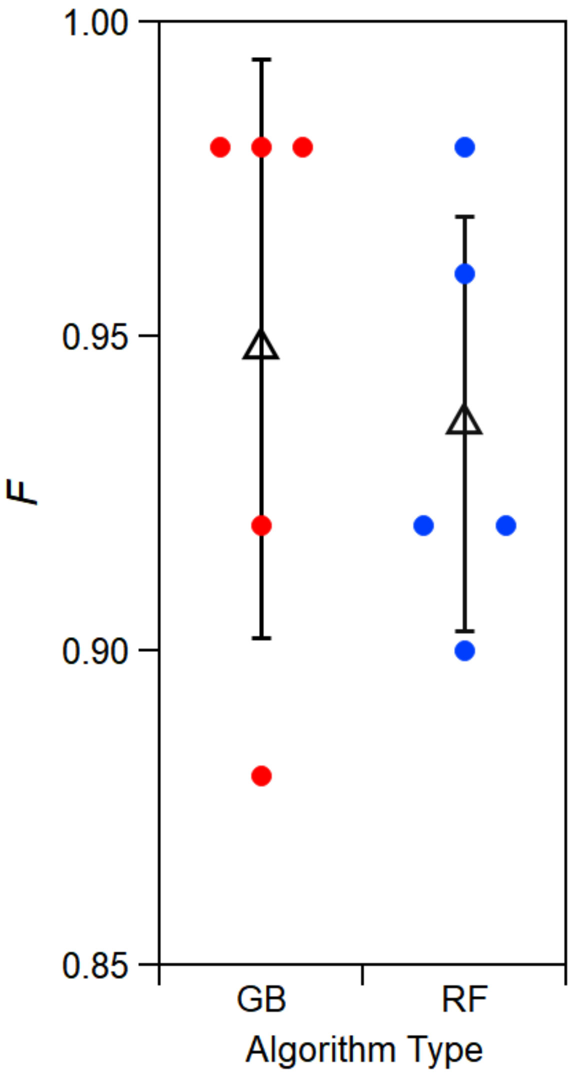


Figure 4.6. Accuracy of GB and RF algorithms summarized after five randomized trials using the 8-species data set. Average accuracy shown as triangle marker. Vertical line shows standard deviation (0.05 for GB, 0.03 for RF). Colored markers show results from individual trials. Identical randomized sets and numerical seeds were used for all trials.

4.7 Random Forest Variable Importance

The relative importance of each portion of each spectrum (770 individual variables accumulated over four emission spectra analyzed at 1 nm resolution) can be determined from the developed RF model as MDG plotted as a function of input variable. Figure 4.7 thus indicates how each emission curve feature can influence the RF algorithm results.

This analysis suggests that the relative importance of the emission spectrum following 280 nm excitation closely follows the pattern of the spectrum itself (Fig. 4.7a). The same is true for the spectrum following 450 nm excitation (Fig. 4.7d), but with overall lower MDG value. The shape of the variable importance curves for the remaining two spectra (Figs. 4.7b and 4.7c) present flatter relationships, in some cases with increasing MDG at tails of the spectrum. The shapes of these curves may imply RF model development misled by noise unfiltered by this method, but also suggests that minor features of the spectra may be important for classification, even if not clearly visible in emission spectra averaged from many individual particles. Variable importance measured before data was noise-filtered (as discussed in Section 2.1) is shown in supplemental Appendix Figures A8 and A9. Particle size variables were input as three independent variables (major and minor axes and aspect ratio of particle size). The MDG values summed for these three size variables (input as 3 of 773 variables after noise filtration) totaled only 1% of the total model MDG for the RF model, whereas integrated values for emission curves correspond to 31%, 46%, 18%, and 4% for 280, 350, 405, and 450 nm excitation sources, respectively.

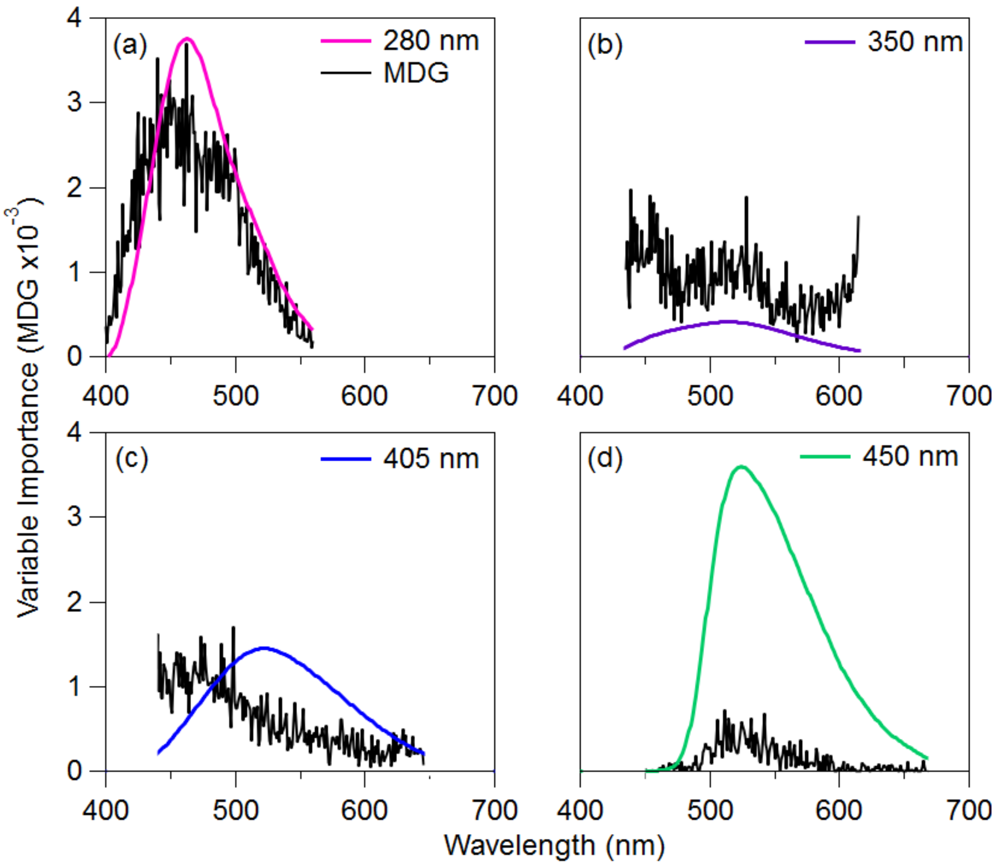


Figure 4.7. Comparison of variable importance for emission spectral following each excitation wavelength for the RF algorithm. Black traces show MDG value. Colored traces show average fluorescence spectra for *Ambrosia trifida* (N=30) shown here as an example.

Pollen analyses are frequently conducted using particle size and shape analysis, even without any additional information such as spectra (e.g. Weber 2010; Bragg 1969; Jones and Newell 1948). We postulated that developing a model that relies only 1% on physical dimensions of the particles would weaken the classification power of the technique. To counteract the under-representation of particle size within the model, each of the three particle size variables was weighted more heavily in order to increase their influence on the model. By inputting each of three size variables as 33 identical columns of data, the total fraction of input size variables was increased to 99 / 869 = 11.4%. The weighting factor was chosen arbitrarily so the observed MDG values of the major and minor diameter variables were on the same order of magnitude as the MDG values for emission spectra (e.g. Fig. 4.8). Weighting particle size increased *F* by a factor of 2.4. Further testing could be conducted on how best to optimize the weighting, but the value utilized was sufficient for the RF model to utilize a mix of particle size and fluorescence information for classification. In this sense, the effect of the scale of weighting used here is less important than its relative effect of arbitrarily increasing particle size importance.

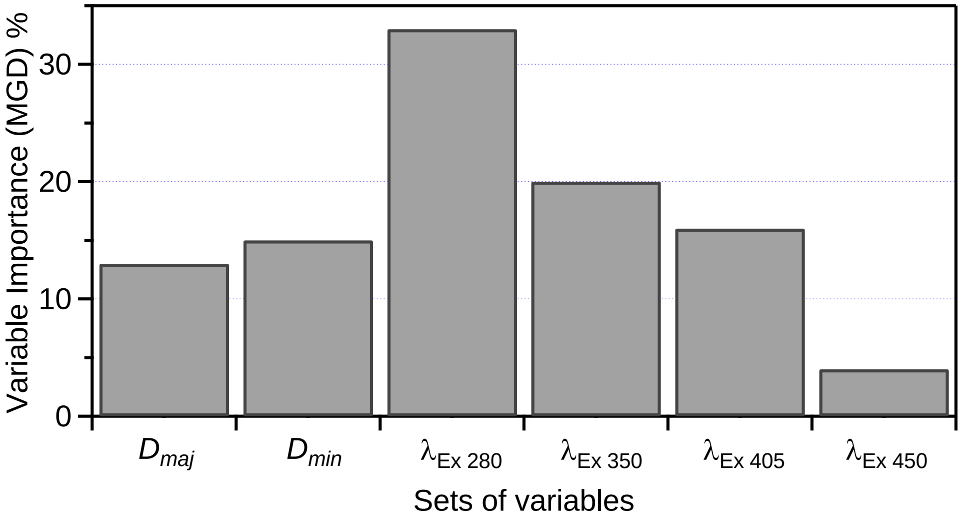


Figure 4.8. Variable importance (MDG) represented as a fraction of total importance. Wavelength of excitation (λ_{Ex}) represents sum of emission variables associated with each optical source. Bars sum to 100%. Particle size aspect ratio removed for visual clarity (showed 0% importance).

The increased importance of particle size after weighting is shown in Supplemental Figure B9. Using the weighted particle size inputs, Figure 4.8 shows that emission spectra following 280 nm excitation exhibit the highest total importance (32% of total MDG), emission spectra following 450 nm excitation are the least important (3%), and the other two emission spectra and particle axes parameters each represent relatively similar influence (15-22%). It is important to note that the values shown in Figure 4.8 are integrated over full curves in Figure 4.7 and thus show identical overall trends.

4.8. Instrument Simplification

In order to further investigate the relative importance of each of the excitation sources, the mean model accuracy (*F*) was calculated after removing different combinations of input variables, each associated with a given excitation source. The purpose of this analysis was to investigate the relative loss of sensor functionality if developed with fewer optical excitation sources, thus producing a less expensive and simpler instrument. All combinations of sources were analyzed (16 in total), corresponding to the use of all four sources and the removal of one, two, or three sources. For each of these cases, particle size variables were not input to the model to compare the changes in model accuracy as it applies to spectral data. The results of the analysis are shown for the test set in Figure 4.9 (training set in Appendix Fig. A10). To simplify discussion, nomenclature here is used such that the 280, 350, 405, and 450 nm sources are labeled as source A, B, C, and D, respectively. For the case using all four excitation sources (case ABCD), the relative accuracy of the model (0.93 training, 0.92 test) was higher than in all cases where at least one source was removed. The test set accuracy for each of four cases where a single source was remove ranged from 0.84 (BCD) to 0.92 (ABC), for the six cases involving two excitation sources from 0.68 (CD) to 0.82 (BC), and for the four cases involving only one source from 0.41 (D) to 0.70 (B). Interestingly, case B showed higher accuracy in both training and trial sets than case BD, implying that the additional input data from the 450 nm source may be confusing model development.

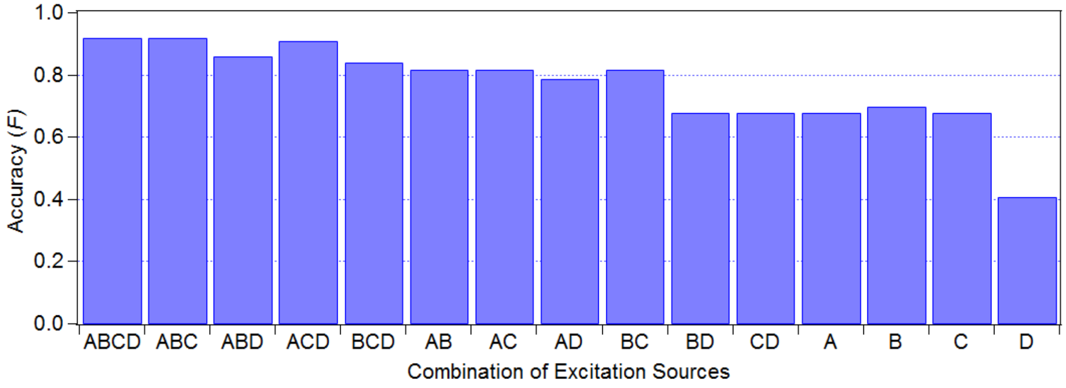


Figure 4.9. Accuracy of the RF algorithm following fifteen combinations of input variables. Excitation sources represented here as (A) 280 nm, (B) 350 nm, (C), 405 nm, and (D) 450 nm. All trials consist of a subset of 25% of the particle spectral data predicted to training models from 75% of the data.

The highest accuracy results from the test set were provided when using all four optical sources (*F* 0.92), which is not surprising. The comparable accuracy of the ABC (450 nm source removed; *F* 0.92) and ACD (350 nm source removed; *F* 0.91) cases suggests here, however, that the relative additional value of either the 350 nm or 450 nm source is marginal toward pollen differentiation. By further simplifying the instrument to use only two sources, the accuracy diminishes somewhat, but all combinations of the 280 nm source (A) plus another source provide relative equal accuracy (0.79 – 0.82). Interestingly, the BC case (350 and 405 nm sources) provided nearly identical accuracy (0.82) to the cases involving the 280 nm source, whereas the remaining two cases (BD and CD) were substantially lower in accuracy. In cases utilizing only one optical source, the relative accuracy diminished still further, but cases involving the 280, 350, and 405 nm sources were nearly identical, whereas the 450 nm source provided clearly the lowest accuracy results.

The relatively poor performance of the 450 nm source, observed in analyses associated with Figures 4.7-4.9 is striking, especially given the ubiquity of emission mode VI in most previously analyzed pollen species (Appendix Figure A6). We originally considered that a reason for the lack of importance to this source was influenced by relatively consistent concentrations of fluorophores (e.g. carotenoid compounds) comprising this mode. Figure 4.5 suggests this is not the case, however, with mean peak emission intensity following 450 nm excitation varying from 2 to 15 arbitrary intensity units. As discussed, the 405 nm source promotes fluorescence from the tails of emission spectra from both the ~350 nm (mode III; phenolics) and ~405 nm (mode V; carotenoids) sources, and thus are expected to be comprised of emission from both sets of fluorophores. The peak height of emission spectra following 405 nm excitation are lower than spectra following 450 nm excitation for seven of the eight pollen species analyzed here, consistent with the spectra that would be expected if the 405 nm spectra were dominated by mode V (peak emission 520 nm) over mode III (peak emission 450 nm). The relatively low importance of the 450 nm source compared to either the 350 nm or 405 nm sources, however, suggests that the 405 nm excitation gains enough information from mode III to reduce the relative additional value of the 450 nm source.

4.7.2 Discussion and Conclusions

Pollen monitoring and forecasting is relatively expensive and time-consuming due to its manual nature. This leads to poor spatial coverage of measurements sites, further leading to models with relatively poor spatial accuracy. We previously presented the development of a single-particle fluorescence spectrometer designed primarily toward inexpensive, portable, and autonomous differentiation of allergenic pollen. The analysis discussed here shows the application of four styles of computational classification to most accurately differentiate between properties of individual particles from eight species of commercially-acquired pollen. We conclude that the GB and RF models provide nearly identical, high accuracy with respect to the pollen species interrogated and that the RF model better optimized cost-benefit with respect to separation accuracy and computational cost.

Fluorescence spectra of pollen and other biological aerosol particle types are relatively broad by physical nature and show relatively similar spectral properties between species, thus it has long been suggested that single-particle differentiation between species would be impossible or challenged by high uncertainty (Huffman et al. 2019; SC Hill et al. 1999). The results here, however, show high levels of separation accuracy using particle size and well-resolved emission spectra acquired from four excitation sources. In these cases, relatively subtle differences in emission intensity associated with many different chemical compounds present in the pollen are likely why separation can be so effective here. This stands in analogy to other methods that separates e.g. species of bacteria based on differences in relative proportion of individual lipid molecule concentrations (e.g.

Madonna et al. 2001). For this reason, separation is improved by the acquisition of relatively high-resolution spectra and is thus improved over single-particle techniques that acquire fluorescence emission data in only 1-3 emission channels or with fewer excitation sources.

The analysis of input variables shown here suggests that the 280 nm excitation source is individually the most important of excitation sources utilized (Figs. 4.8 and 4.9), but that its importance is somewhat diminished when comparing results after removing individual sources from the analysis. It is clear that developing an instrument with all four excitation sources can provide high classification accuracy. Altering the design to utilize three or less source may be an attractive solution, however, to reduce cost and complexity. Of the six two-source cases, three cases (AB, AC, BC) performed approximately equally well. Two of those utilized the 280 nm source, which is not surprising given its overall importance (Fig. 4.8). More surprising was that two of these three cases utilized the 405 nm source, which performed with lower overall integrated importance (Fig. 4.8). In cases where only a single optical source was utilized, the 280, 350, and 405 nm sources performed with similar mean accuracy. From a practical perspective, it is advantageous to choose sources that minimize cost and maximize longevity (e.g. robust, high cumulative operation time) and that provide enough output power density that promote emission spectra sufficiently intense to allow shorter image integration times. In this context, the 280 nm source is comparatively expensive and provides the weakest output power (0.33 mW compared to 4.5 – 50 mW for other sources). The weak output power leads to longer exposure times (~3 minutes) for images of particle sets compared to the other three sources (3-50 seconds). For these reasons, a combination of the 350 nm and 405 nm sources (BC) may be ideal for a small detection platform, because demonstrated mean accuracy is high, and both cost and practical convenience are advantageous. In comparison, the addition of the 450 nm source to the 350 and 405 nm sources (BCD) adds only a 2% improvement on the classification, which may not result in sufficient accuracy gain relative to the additional material cost.

The results shown are important not only toward the future development and application of the instrument discussed specifically here, but more broadly to emerging classes of instrumentation that acquire complex data (spectral or otherwise) with many variables toward the purpose of single-bioparticle differentiation. In particular, the analysis of the importance of individual input parameters within a spectrum, integrated groups of variables, and the relative differences in model accuracy after removing instrument components can provide context for development and testing of emerging particle spectrometers.

The primary goal of the single-particle fluorescence spectrometer discussed here is toward the detection of allergenic pollen species in approximately real-time, so as to contribute to the areas of missing data and to improve spatial accuracy in pollen forecasting models. In this context, the scientific application of the measurement does not require species-level identification of all airborne pollen, but rather the differentiation of the species of highly allergenic pollen that dominate the public health response. Thus, to lower detection and analysis requirements, collection of lower resolution spectra may be sufficient for adequate prediction. Future analysis will thus investigate the trade-offs by either collecting spectra at lower resolution or by parameterizing spectra into fewer input variables (e.g. as averaged intensity in the eight fluorophore modes discussed here). For now, however, the computational requirements of the RF model are sufficiently low that it is not expected to be a limiting factor in the particle analysis, and preliminary work suggests that the subtle nuances in the high-resolution spectra as collected contribute positively to accurate pollen differentiation

Chapter 5: Pollen Classification with a Recently Developed Fluorescence Spectrometer

The information presented in this chapter is in preparation for submission for peer review.

5.1 Introduction

Pollination mechanisms can be separated into two groups: abiotic, pollination based on environmental factors, and biotic, pollination based on pollinator symbioticism. The majority of gymnosperms, grasses, sedges, and rushes are all anemophilous, or wind-pollinated species, as well as some other trees such as oak, walnut, or chestnut (Ackerman 2000). In order to be transported through the atmosphere, these types of pollen tend to be smaller and less, or non-, sticky compared with pollinator-transported pollen (Ackerman 2000). Anemophilous plants, such as those from the genus *Pinus*, have been seen to produce 100,000 pollen grains per individual anther (Molina et al. 1996). Similarly, tree types with longer anthers have been seen to produce larger numbers of pollen grains (Molina et al. 1996). Anemophilous pollen contributes to the majority of those particles measured in the atmosphere, up to 98% of the pollen sum measured (Kasprzyk 2004; Mullins and Emberlin 1997), though large concentrations of entomophilous pollen have been previously seen in atmospheric samples in conjunction with high wind speeds (Dua and Shivpuri 1962).

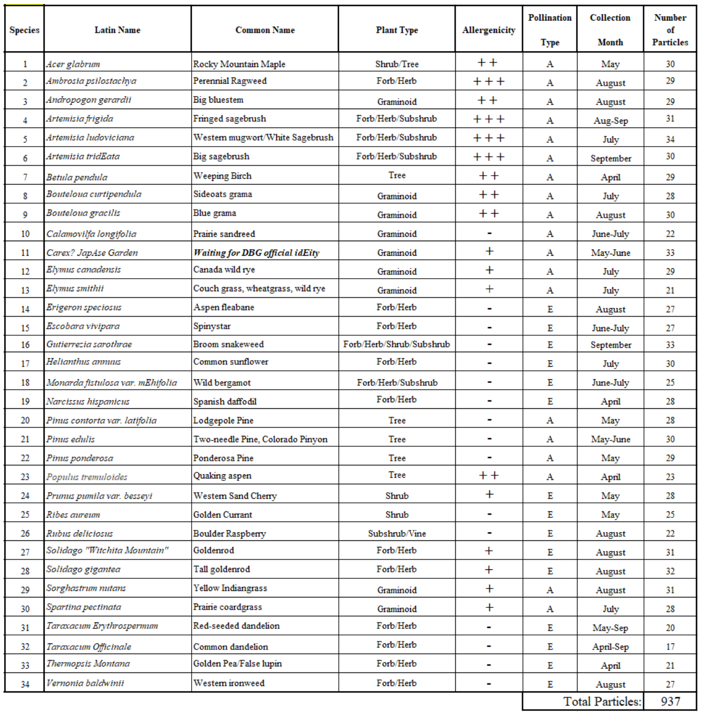
Previous single-particle studies on the fluorescence of pollen have also shown variation in individual grains of the same species (O'Connor et al. 2014b; O'Connor et al. 2011), let alone between species. Single-particle data gives a much more reliable view of the distribution within species, as well as between, due to the statistical variation between individual particles. Bulk sampling can give good information about the differences between species, as well as any differences seen by regional or local variations in soil quality, weather, and other factors which may be directly related to pollen viability at the time of measurement (Nyomora et al. 2019; Khatun and Flowers 2007). Certain pollen types, such as grass pollens, have also shown to contain chlorophyll *a* that is fluorescently available to instrumentation (D. O'Connor et al. 2014b).

5.2 Methods

5.2.1 Sample Collection and Treatment

A total of 34 species of pollen were collected between May 2018 and May 2019 from plants growing at either the Denver Botanic Gardens or the University of Denver arboretum campus (Table 5.1). Samples were chosen to select a range of plant types, allergenicity levels, and pollination mechanisms. In some cases, plants were chosen as examples of those indigenous to the region. Two different sampling methods were utilized, depending on the pollination mechanism of plant. Anemophilous pollen samples were collected by gently shaking the flower onto a glass slide. Entomophilous pollen samples were collected either by shaking the flower onto a glass slide or by transferring pollen grains from the stamen of the flower using a small needle and immediately transferring pollen grains to a slide.

Sample slides were inserted into plastic slide holders for transport and storage. In most cases, multiple slides were collected for a single species, and different species were always stored separately. Sample holders were closed and stored in a refrigerator (~4 ^{o}C) within 1-3 hours of collection. Imaging and spectral analysis was performed within a maximum of five days after collection. Between 20 and 30 particles were analyzed for each individual species. During imaging analysis, pollen grains were qualitatively assessed to ensure that the morphology and size of the pollen approximately matched images in a pollen database that contained scanning electron microscope and optical micrograph images (Weber and Ulrich 2017).



Species	Latin Name	Common Name	Plant Type	Allergenicity	Pollination Type	Collection Month	Number of Particles
1	*Acer glabrum*	Rocky Mountain Maple	Shrub/Tree	++	A	May	30
2	*Ambrosia psilostachya*	Perennial Ragweed	Forb/Herb	+++	A	August	29
3	*Andropogon gerardii*	Big bluestem	Graminoid	++	A	August	29
4	*Artemisia frigida*	Fringed sagebrush	Forb/Herb/Subshrub	+++	A	Aug-Sep	31
5	*Artemisia ludoviciana*	Western mugwort/White Sagebrush	Forb/Herb/Subshrub	+++	A	July	34
6	*Artemisia tridEata*	Big sagebrush	Forb/Herb/Subshrub	+++	A	September	30
7	*Betula pendula*	Weeping Birch	Tree	++	A	April	29
8	*Bouteloua curtipendula*	Sideoats grama	Graminoid	++	A	July	28
9	*Bouteloua gracilis*	Blue grama	Graminoid	++	A	August	30
10	*Calamovilfa longifolia*	Prairie sandreed	Graminoid	-	A	June-July	22
11	*Carex? JapAse Garden*	***Waiting for DBG official idEity***	Graminoid	+	A	May-June	33
12	*Elymus canadensis*	Canada wild rye	Graminoid	+	A	July	29
13	*Elymus smithii*	Couch grass, wheatgrass, wild rye	Graminoid	+	A	July	21
14	*Erigeron speciosus*	Aspen fleabane	Forb/Herb	-	E	August	27
15	*Escobara vivipara*	Spinystar	Forb/Herb	-	E	June-July	27
16	*Gutierrezia sarothrae*	Broom snakeweed	Forb/Herb/Shrub/Subshrub	-	E	September	33
17	*Helianthus annuus*	Common sunflower	Forb/Herb	-	E	July	30
18	*Monarda fistulosa var. mEhifolia*	Wild bergamot	Forb/Herb/Subshrub	-	E	June-July	25
19	*Narcissus hispanicus*	Spanish daffodil	Forb/Herb	-	E	April	28
20	*Pinus contorta var. latifolia*	Lodgepole Pine	Tree	-	A	May	28
21	*Pinus edulis*	Two-needle Pine, Colorado Pinyon	Tree	-	A	May-June	30
22	*Pinus ponderosa*	Ponderosa Pine	Tree	-	A	May	29
23	*Populus tremuloides*	Quaking aspen	Tree	++	A	April	23
24	*Prunus pumila var. besseyi*	Western Sand Cherry	Shrub	+	E	May	28
25	*Ribes aureum*	Golden Currant	Shrub	-	E	May	25
26	*Rubus deliciosus*	Boulder Raspberry	Subshrub/Vine	-	E	August	22
27	*Solidago "Witchita Mountain"*	Goldenrod	Forb/Herb	+	E	August	31
28	*Solidago gigantea*	Tall goldenrod	Forb/Herb	+	E	August	32
29	*Sorghastrum nutans*	Yellow Indiangrass	Graminoid	+	A	August	31
30	*Spartina pectinata*	Prairie coardgrass	Graminoid	+	A	July	28
31	*Taraxacum Erythrospermum*	Red-seeded dandelion	Forb/Herb	-	E	May-Sep	20
32	*Taraxacum Officinale*	Common dandelion	Forb/Herb	-	E	April-Sep	17
33	*Thermopsis Montana*	Golden Pea/False lupin	Forb/Herb	-	E	April	21
34	*Vernonia baldwinii*	Western ironweed	Forb/Herb	-	E	August	27
						Total Particles:	937

Table 5.1. The 34 species collected over the 2018-2019 pollination season. Common name, plant type per USDA classification, allergenicity level: low (0) to severe (3), pollination mechanism type, and the particle number collected and subsequently analyzed for each individual species.

5.2.2 Pollen Species Categorization

To test the ability of the instrument and analysis technique to separate pollen, species were chosen in part to represent a variety of physical and biological properties. Categorization of species was considered with respect to four separate methods of organization: (i) plant type, (ii) allergenicity level, (iii) pollination mechanism, and (iv) month of sample collection. Each plant type was categorized into one of five groups according to the USDA definition (USDA 2016): forb (flower herbs), graminoid (grasses), shrubs, subshrubs, or trees.. Allergenicity level was categorized into one of four groups (0, none; 1, mild; 2, average; 3, severe) as defined by IMS Health (http://www.pollenlibrary.com). Pollination mechanism was defined as either anemophilous or entomophilous. All category determinations are listed for each species in Table 5.1. Month of sample collection was further organized into four longer sampling periods to account for the fact that pollination can extend beyond an individual month for a given species. Sampling period were grouped as: P1 (April, May, June), P2 (May, June, July), P3 (June, July, August), and P4 (July, August, September). In this way, samples collected in a given month are present in multiple sampling periods, as a sort of rolling average, to examine the quality of pollen separation in each of these cases.

5.3 Pollen Classification Results

5.3.1 Overview of Pollen Data

A total of 34 species of pollen were analyzed, representing diversity of collection month, plant type, allergenicity level, and pollination mechanism. The species collected also exhibited a wide range of particle size (24 – 86 μm), size aspect ratio (1.02 – 2.30), and spectral characteristics. Results from eight species were chosen to highlight the range of observed properties and model prediction qualities, e.g. the species that were most accurately (Figs 5.1a-d; top four rows) and least accurately (Figs 5.1e-h) tablsubscripts for these represent their positioning in the overall 34 species data set, which is shown in Appendix Figures B3-7.

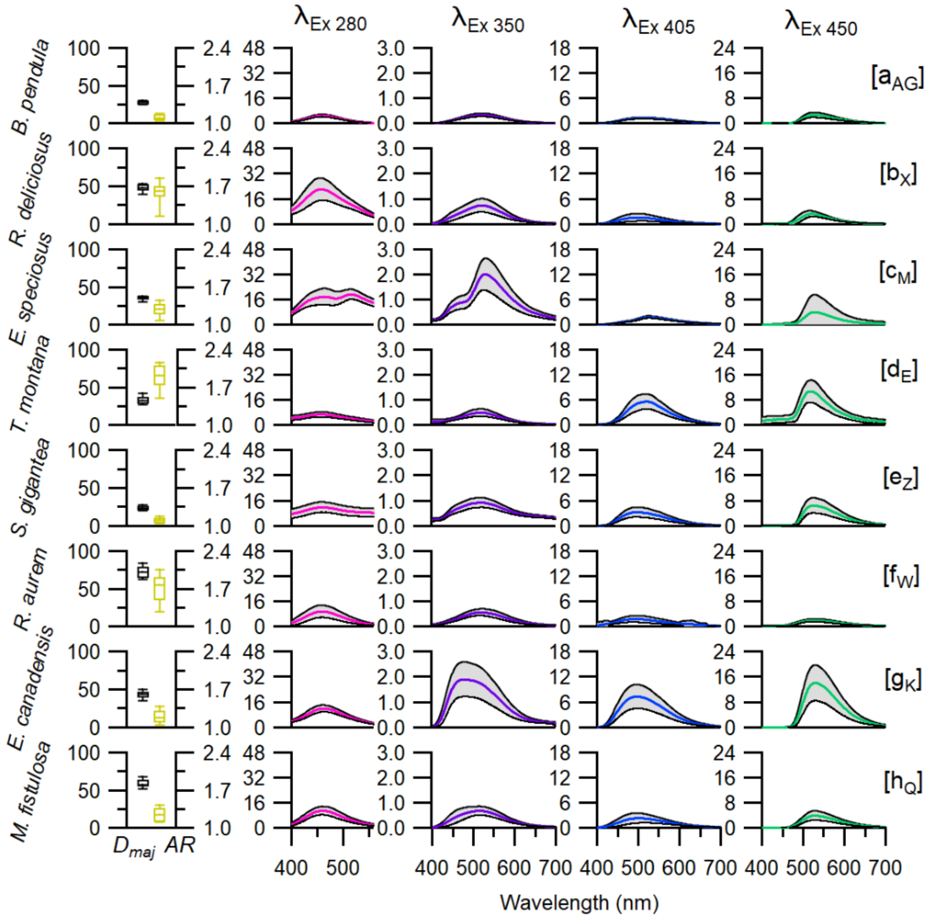


Figure 5.1. A selection of various pollen collected throughout the 2018-2019 seasons. Species are labeled here as A-H, with their subscripts listing their position in the full 34 species data set. Box and whisker plots for the major size axis (black) and aspect ratio (yellow) are shown in the left column of the plot. This is followed by the average emission curves for the 280, 350, 405, and 450 nm excitation sources, from left to right. Grey bounds are drawn around each curve, representative of the deviation in those species' sets. Each emission column has the intensity conserved to make visual comparisons easier.

Each excitation wavelength probes a different set of fluorophore modes present in the pollen species. The emission modes following 405 and 450 nm excitation exhibit similar signals all species shown here per source. Generally, there is a single mode present near 500 nm following excitation source 405, (V); and one near 520 nm following 450 nm excitation, (VI), corresponding to the fluorescence e.g. carotenoids. There are small variations in maximum peak positioning (e.g. *E. speciosus*; species c_M and *E. canadensis*; species g_K emission mode following 405 nm excitation being shifted near 520 nm than 500 nm), though the main differences between individual species is the peak intensity of the fluorophore modes seen. One case, *E. speciosus* (species c_M), shows an additional shoulder mode after 405 nm excitation, at 475 nm emission (similar to the 350 nm excitation), corresponding to emission mode (III) e.g. phenolic compounds. This may be due to the 405 nm source probing *a minimum* between two major fluorophore groups previously reported as observed previously (Pöhlker et al. 2013), consisting of e.g. phenolic and carotenoid signals, and these varying concentrations are shifting where the main peak is detected. This effect is seen where these fluorophore concentrations are seen varying in emission from the 350 nm source as well (*E canadensis*; species g_K has higher mode (II) peak; *H. annus*; species P has higher mode (IV) peak), where the single 405 nm source emission peak position can be seen varying in a similar way. The modes from 450 nm, in contrast, only show mode (VI) for e.g. carotenoids, with varying intensities. These two excitation modes were shown to previously be less important in the RF model development (Swanson and Huffman 2019), which is shown through the lack of variability in the emission modes for these two sources. Notably, though previous studies have noted that certain pollen species may exhibit, such as grass pollen (O'Connor et al. 2011), no grass pollen species in this data set contained any chlorophyll *a* that was available to probe by this instrument. This leaves out emission modes (VII) and (VIII) completely.

Emission signals from the 350 and 280 nm sources show wide variability when compared to the previous two sources. Emission from the 280 nm source largely shows an emission mode (I) around 430-460 nm, corresponding to e.g. phenolic structures, with varying intensity and shoulder features. The features from this emission mode tends to revolve around the positioning of the peak itself (430 to 460 nm), rather than several different peaks. Still, there is a mode not previously described explicitly in previous studies, likely from e.g. carotenoids, near 520 nm in *E. speciosus*; species c_M, which shows up as a clearly distinct second peak in the 280 nm emission spectra. *S. gigantea*; species e_Z, and *T. montana*; species d_E, appear to show a longer shoulder in that area as well, which is not seen in the other six species. Emission from the 350 nm source shows the highest variability across all sources. Two key modes, which correspond to varying degrees of e.g. phenolics (II) and e.g. carotenoids (IV) present in an individual pollen grain, are commonly seen. In some species, this results in one large peak area with a shoulder (*B. pendula, R. deliciosus, T. montana, S. gigantea*, *R. aurem*) while others have shoulders or peaks that are clearly distinct from the main peak (*E. speciosus, E. canadensis, M. fistulosa*). The emission from the 350 nm source exhibiting the highest qualitative variability is consistent with previous data showing RF models are developed heavily off of this excitation source, as well as off the 280 nm source (Swanson and Huffman 2019).

A range of sizes and aspect ratios are seen in this data set, as well. These differences are visually present in Fig. 5.1 and Appendix Fig. B3-7. Many of these differences include species with extremely narrow size distributions (i.e. *A. frigida*; species D, *G. sarothrae*; species O, *S. gigantae*; species Z). Species of the same genus (Artemisia; species D-F, Pinus; species S-U, and Solidago; species Y and Z) have remarkably similar size characteristics, and generally similar aspect ratio characteristics as well. In these cases, different spectral characteristics between the individual species seem to be the driver in classifications to each species, as opposed to size and morphology. In comparisons between different genus, such as *Betula* (AG) and *Ribes* (W), where size characteristics don't overlap at all, though there is some overlap in spectral similarities, ensuring size is able to drive these separations is important. Considering no misclassifications between these two groups, it helps justify the previous implementing of artificial size weighting for this model development (Swanson and Huffman 2019).

5.3.2 Differentiation of Entire Data Set

The first classification scenario performed was the species level classification of the entire, 34-species data set. The results of this classification scenario are shown in Figure 5.2, which shows the ratio of correctly and incorrectly classified particles for each individual species, and the mean of the entire classification set (*F* value 0.89). The confusion matrix associated with this classification can be seen in Table 5.2. The direct misclassifications for each individual species can be seen in Table 5.2. Half of the species were classified to an *F* value of 0.90 or greater, and only a few species (*M. fistulosa*, 0.77; *R. aurem*, 0.81; *E. canadensis*, 0.81) showed high levels of misclassification, though none reported low *F* values than 0.77. *E. canadensis* is misclassified to four total species, though three of the six misclassified particles were classified as *V. baldwinii*. Interestingly, *M. fistulosa* was misclassified evenly to *E. vivipara* and *R. aurem*, the latter being one of the more misclassified species. The major diameter characteristics for *M. fistulosa* match extremely closely to both species, while the spectral characteristics are similar to *R. aurem* and are also similar to *E. vivipara* (though less so with the emission from the 280 nm source), and the aspect ratio is extremely similar to *E. vivipara*. *R. aurem* was also similarly misclassified twice to *E. vivipara*, likely for similar reasons. Two other interesting examples are *P. ponderosa*, which has misclassifications of a single particle each to the other two *Pinus* species present. This type of intra-genus misclassification also happened for *Solidago,* in which they exchange a total of three misclassified particles. These examples indicate that there may be some appreciable crossover between individual species of the same genus, though larger sample sizes may be needed to confirm this.

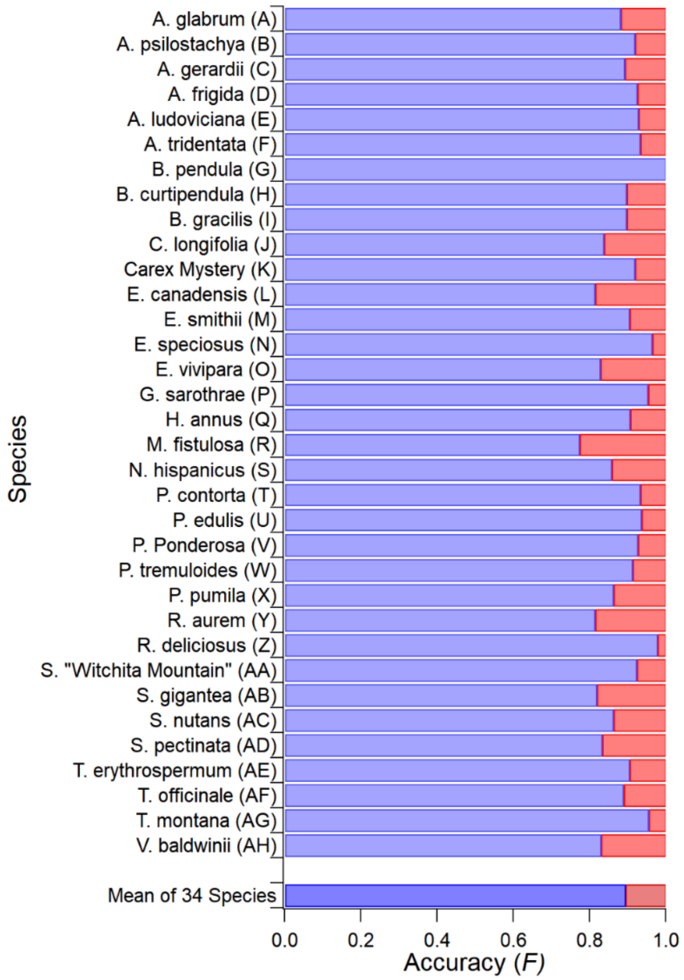
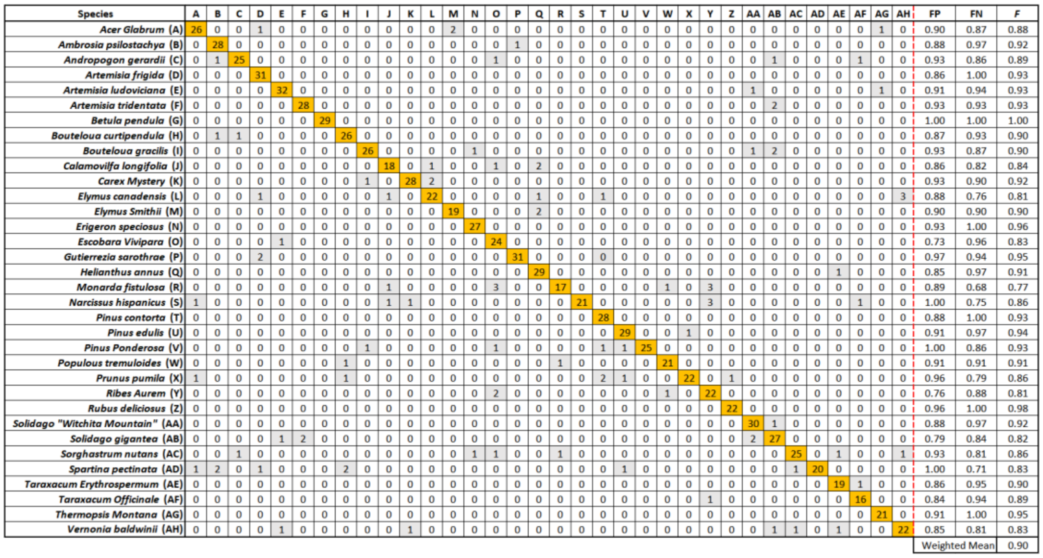


Figure 5.2. The accuracy and misclassification for each individual species for the collections from the Denver Botanic Gardens and University of Denver. The blue bar refers to correct classifications, while the red bar refers to misclassifications, adding up to 1. A total of 933 particles were classified here, with the overall data set being classified at 89% accuracy.

For higher accuracy trials, *E. speciosus* is shown in Figure 5.1 as being qualitatively extremely different than the others, as evidenced by the fluorescent modes (II), (IV), and (III), as well as the shoulder for (V) present in this species, corresponding to the phenolic and carotenoid peaks. This indicates that the spectral and size differences in these 34 species, despite there being some cases of large distributions in spectral intensity and positioning, can lead to a classification accuracy of 0.89.



Species	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD	AE	AF	AG	AH	FP	FN	*F*
Acer Glabrum (A)	26	0	0	1	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0.90	0.87	0.88
Ambrosia psilostachya (B)	0	28	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.88	0.97	0.92
Andropogon gerardii (C)	0	1	25	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0.93	0.86	0.89
Artemisia frigida (D)	0	0	0	31	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.86	1.00	0.93
Artemisia ludoviciana (E)	0	0	0	0	32	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0.91	0.94	0.93
Artemisia tridentata (F)	0	0	0	0	0	28	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0.93	0.93	0.93
Betula pendula (G)	0	0	0	0	0	0	29	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1.00	1.00	1.00
Bouteloua curtipendula (H)	0	1	1	0	0	0	0	26	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.87	0.93	0.90
Bouteloua gracilis (I)	0	0	0	0	0	0	0	0	26	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	2	0	0	0	0	0	0	0.93	0.87	0.90
Calamovilfa longifolia (J)	0	0	0	0	0	0	0	0	0	18	0	1	0	0	1	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.86	0.82	0.84
Carex Mystery (K)	0	0	0	0	0	0	0	0	1	0	28	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.93	0.90	0.92
Elymus canadensis (L)	0	0	0	1	0	0	0	0	0	1	0	22	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0.88	0.76	0.81
Elymus Smithii (M)	0	0	0	0	0	0	0	0	0	0	0	0	19	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.90	0.90	0.90
Erigeron speciosus (N)	0	0	0	0	0	0	0	0	0	0	0	0	0	27	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.93	1.00	0.96
Escobara Vivipara (O)	0	0	0	0	1	0	0	0	0	0	0	0	0	0	24	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.73	0.96	0.83
Gutierrezia sarothrae (P)	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	31	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.97	0.94	0.95
Helianthus annus (Q)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	29	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0.85	0.97	0.91
Monarda fistulosa (R)	0	0	0	0	0	0	0	0	0	1	0	0	0	0	3	0	0	17	0	0	0	0	1	0	3	0	0	0	0	0	0	0	0	0	0.89	0.68	0.77
Narcissus hispanicus (S)	1	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	21	0	0	0	0	0	3	0	0	0	0	0	0	1	0	0	1.00	0.75	0.86
Pinus contorta (T)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	28	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.88	1.00	0.93
Pinus edulis (U)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	29	0	0	1	0	0	0	0	0	0	0	0	0	0	0.91	0.97	0.94
Pinus Ponderosa (V)	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	1	1	25	0	0	0	0	0	0	0	0	0	0	0	0	1.00	0.86	0.93
Populous tremuloides (W)	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	21	0	0	0	0	0	0	0	0	0	0	0	0.91	0.91	0.91
Prunus pumila (X)	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	2	1	0	0	22	0	1	0	0	0	0	0	0	0	0	0.96	0.79	0.86
Ribes Aurem (Y)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	1	0	22	0	0	0	0	0	0	0	0	0	0.76	0.88	0.81
Rubus deliciosus (Z)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	22	0	0	0	0	0	0	0	0	0.96	1.00	0.98
Solidago "Witchita Mountain" (AA)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	30	1	0	0	0	0	0	0	0.88	0.97	0.92
Solidago gigantea (AB)	0	0	0	0	1	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	27	0	0	0	0	0	0	0.79	0.84	0.82
Sorghastrum nutans (AC)	0	0	1	0	0	0	0	0	0	0	0	0	0	1	1	0	0	1	0	0	0	0	0	0	0	0	0	0	25	0	1	0	0	1	0.93	0.81	0.86
Spartina pectinata (AD)	1	2	0	1	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	20	0	0	0	0	1.00	0.71	0.83
Taraxacum Erythrospermum (AE)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	19	1	0	0	0.86	0.95	0.90
Taraxacum Officinale (AF)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	16	0	0	0.84	0.94	0.89
Thermopsis Montana (AG)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	21	0	0.91	1.00	0.95
Vernonia baldwinii (AH)	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	1	0	0	22	0.85	0.81	0.83
																																			Weighted Mean		0.90

Table 5.2. Confusion matrix of the entire, 34-species data set classified to the species level. On the left side of the dashed red line is the matrix itself, with correct classifications coded in orange, and misclassifications coded in grey. To the right of the red dashed line is the ratio of false positives (FP), false negatives (FN), and the overall misclassification vector (*F*)

Though separation at the species level can be important, many of these species were collected in different months of the year, a result of them having different pollination seasons. Many species pollinating in early March and April, usually tree species, will not be pollinating simultaneously with ragweed or other pollinating subshrubs like blue sagebrush. So, though successful, species-level separation of all pollen types is not of high importance. Moving forward, the goal is to identify and classify specific, important species that are common allergens for a certain area or time period. In this case, the focus was solely on pollen frequently found in the Colorado front range area.

5.3.3 Analysis Separation by Collection Period and Plant Metadata

Overview

Species collected at the Denver Botanic Gardens were largely constrained by both the availability of individual species, as well as the timeframe of possibly collection. Utilizing the sampling windows described previously, the pollen species being analyzed can be restricted to a series of classifications where there is reasonable expectation that the species being compared may be pollinating contemporaneously. Not only can this be done at the species level, but the pollen can be classified based on metadata associated with those species. As an example, all of the particles from April to June were listed in window P1. In addition to the seasonal window, the type of plant and allergenicity level of the plant can also be utilized to narrow these scenarios, which can be seen in Figure 5.3.

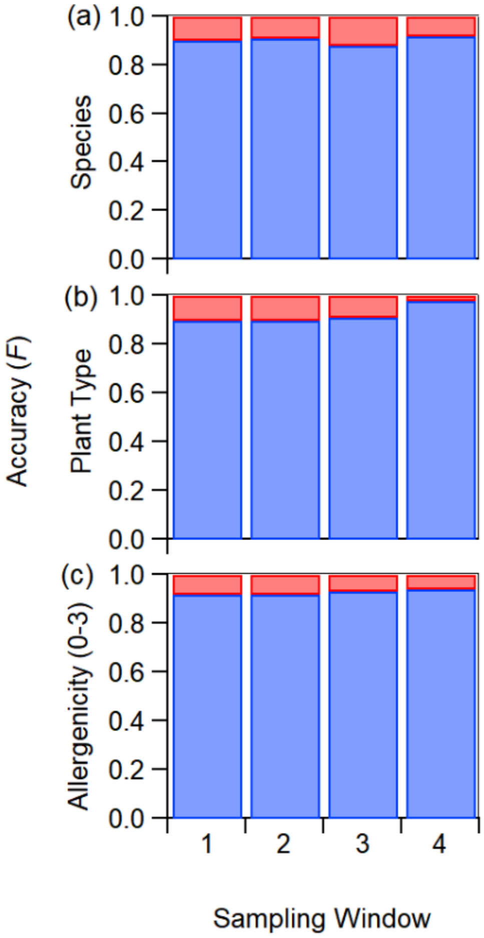


Figure 5.3. Accuracy (*F*) associated with each sampling window, with the fractional amount of correctly classified particles in blue, and the incorrectly classified particles listed in red. These are separated by USDA-listed type, allergenicity level (0-3) and the species level. The entire trial's average error is shown here, not individual types or species.

<u>*Classification by Species*</u>

In comparison with the entire 34-species data set, the species level comparisons were also performed in the four sampling windows, which is shown in Figure 5.3a. Each sampling window at the species level were extremely similar in *F* value to the overall 34-species scenario, all four being within 0.3 of the full set. The lowest accuracy sets were windows P1 and P3, with *F* values of 0.90, which was comprised of 12 species and 309 individual particles. The highest accuracy scenario was window P4, with an *F* value of 0.92, which contained 19 species and 537 total particles. Comparisons of where the species were misclassified can be seen in Appendix Table B2 for one of the lowest accuracy windows (P3) and Appendix Table B3 for the highest accuracy set (P4). Despite P4 containing more species in total, 19, the number of misclassifications more than 1 particle to or from the same species was lower than P3, which had 12 species. For window P3, there were several species that seemed to be particularly problematic. Species C, G, and J in particular, all had 5 or 6 particles misclassified to their respective clusters, which contributed to 44% of the misclassified particles in total. The differences in accuracy between P3 and P4 can easily be seen when examining the species level *F* values, of which 6 of 12 and 4 of 19 were under 0.90, respectively.

<u>*Classification by Plant Type*</u>

The sampling scenarios that are classified by plant type can be seen in Figure 5.3b. In contrast to the species level, all four showed higher accuracy than the total full scenario, showing *F* values at 0.1-1.0 higher. Of note is window P4, which was classified at an *F* value 0.98. A comparison of the lowest accuracy trial (P1) and the highest accuracy trial (P4) can be seen in Appendix Tables B4 and B5, respectively. Though the species count for this were comparable between both windows, 17 for P1 and 19 for P4, the first window contained three more plant type categories than the latter. These differences likely contributed heavily to the larger accuracy difference in the plant type trials, which were the highest, since the plant types for this window P4 are varied the least, having two types (forb and graminoid) present. This contrasts with one of the lower accuracy trials, window P1, which have five total plant types (forb, graminoid, shrub, subshrub, tree). Interestingly, in the P1 trials the graminoids and forbs had some crossover misclassifications between one another, accounting for 16% of the misclassifications in this sampling window, corresponding to 1.6% of the total error. The P4 window here only showed 2.0% total error, which does not seem inconsistent with P1's misclassifications between these two classes, despite being comprised of mostly different species. Still, the possibility remains that if more plant type classes were added, some of these particles would be misclassified differently, as the threshold for assigning is based on similarity.

Classification by Allergenicity

The accuracy of grouping the pollen species by allergenicity (0-3) is marginally better as well, with an *F* value of 0.92-0.94. Window P4 showed the highest accuracy for these trials, like the species and plant type level. Similarly, window P1 showed the lowest accuracy, at an *F* value of 0.92. The confusion matrix comparisons of these can be seen for P1 in Appendix Table B6 and P4 in Appendix Table B7. Window P1 shows many particles being misclassified to the allergenicity level of "none" from both the "mild" and "moderate" groups, which accounts for the majority of misclassifications (91%) for this set. In fact, more "mild" particles were incorrectly classified to the "none" category than were present as correct classifications, or misclassifications to "moderate," leading to the false-negative rate of 0.32 and total *F* value for “mild” of 0.49. It is worth noting that the number of particles for this class only reached 28, and thus a small number of misclassified particles will lead to high error. There was also no particle classified as “severe” allergens in this window. Contrary to this, window P4 does not see any clusters with such drastic effects and the total number of misclassified particles never leads to a FP or FN vector value below 0.94, indicating high accuracy across all classes.

Overview of Reduced Sampling Windows

These improvements in accuracy can be utilized to improve the data classification for these pollen particles collected over the course of the year. It is important to note that even marginal improvements in accuracy are good. Coupled with higher accuracy, reducing the number of classes analyzed simultaneously will subsequently lower the computational burden needed for analysis. Considering the goal of this technique is an inexpensive, portable, autonomous platform, it is important that complex processes are limited to keep costs low. For all data classes, grouping by allergenicity boasted the largest accuracy increase out of any data grouping on average. This is important, as the prospect of being able to report numbers or atmospheric concentrations of moderate to severely allergenic pollen to the public. Still, sampling window P4 in the plant type grouping showed the highest accuracy of any trial.

5.4. Pollen Importance Models for Instrument Simplification

How the RF algorithm treats the data can be indirectly viewed through the usage of variable importance as well as source reduction. The importance for the entire 34-species set can be seen in Appendic Figure C8, which shows the relative importance for six blocks of data (two size; four spectral). This shows that the 280 and 350 nm sources are of extreme importance in the models, accounting for 56% of the developed RF model. The two sizing parameters, major and minor diameter, account for 23% of the total importance. This is interesting since these two single data points are more influential than the combination of two whole spectral sources (405 and 450 nm; 19%). These importance difference indicate that the size parameters themselves may contain more differential information among pollen species than the two longer wavelength sources. This is consistent with previously published information on importance values of these RF models with commercial species (Swanson and Huffman 2019), though the 450 nm source appears to be nearly twice as important (4% to 7%), and the 405 nm source loses some importance (17% to 12%). These differences could be due to differences or similarities introduced by increasing the number of species analyzed from 8 to 34.

To examine the effect individual sources may have in the overall classifications, a series of source reduction trials are shown in Figure 5.4. For this, 75% of the data was utilized as a training set, and 25% of the data was treated as unknowns to be predicted. Size parameters were left out of the trials, and sources were systematically removed. This is analogous to a previous study with commercial pollen and far less species (Swanson and Huffman 2019). By comparison, this showed much less accuracy in unknown predictions, which generally corresponds to the reduction in accuracy among testing sets as well. Generally, a larger combination of the four sources showed a higher prediction accuracy, with the combination of four sources having the highest *F* value (0.65). The three combinations of three sources had *F* values from 0.60-0.63, with the combination of

BCD (350, 405, and 450 nm) showing the lowest accuracy. The two-source combinations were more varied, with *F* values between 0.45 (CD) and 0.58 (AB). This makes sense, as the AB combination showed the highest importance in the previous section, while the CD combination had a combined importance less than the size itself. Curiously, CD had less accuracy than A by itself, and had an *F* value of only 0.01 higher than B, showing that the combination of 405/450 nm sources may be giving information comparable to both single 280 and 350 nm sources. This is important considering this involves the development of an inexpensive, portable pollen detection platform. The potential removal of one or two of these sources can help make this platform viable.

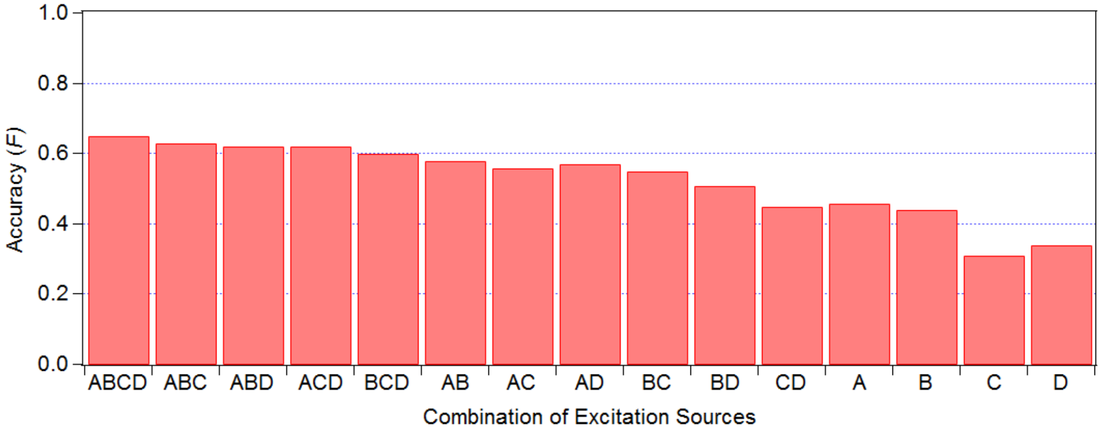


Figure 5.4. Source reduction trials with the entire pollen collection set. Sizing parameters were fully removed, and spectral excitation information was used. Each letter corresponds to an individual source (A – 280; B – 350; C – 405; D – 450). A combination of letters corresponds to multiple sources present in the trial.

5.5 Selected Species Comparisons

5.5.1 Commercial vs. Fresh; Species Comparisons

Two species were tested as crossover species between fresh and commercially purchased types: *Taraxacum officinale* and *Betula pendula*. This is shown for *T.*

officinale in Figure 5.5, which compares commercially purchased pollen from Bonapol (Czech Republic) and freshly collected pollen from the University of Denver campus (Denver, USA). Very clear qualitative differences are seen in these spectral averages. Counterintuitively, three of the four emission curves are higher for the commercially purchased pollen. Each of these three curves exhibit single large peaks, representing fluorophore modes I, V, and VI, while the 350 nm source produces an emission curve equally between fluorophore modes II and IV. All four excitations for the fresh sample show emission modes at 520 nm, corresponding to IV, V, and VI, as well as an additional mode that shows up from 280 nm excitation. All three fluorophore modes associated with 450 nm emission (280, 350, 405 nm excitation) are present in the fresh sample as well. Oddly, the fresh and commercial samples both show roughly the same response of fluorophore mode II, despite drastically different responses otherwise. The difference in collection mechanisms between the two types may be responsible here, as the commercial pollen was reported to have been defatted to remove pollenkitt on the surface of the pollen. Since carotenoid and lipid structures are present in the exine and pollenkitt, the fresh samples having modes IV, V, and VI seems reasonable. Differences could also have been induced by long storage times (1-1.5 years for the commercial sets), as well as any possible degradation in fresh pollen after removal from the stamen. *T. officinale* is an entemophilous pollen type and is unlikely to be seen in atmospheric samples. Anemophilous pollen is also less likely to contain pollenkitt, or at the least contain far less pollenkitt on the surface (Hesse 1984).

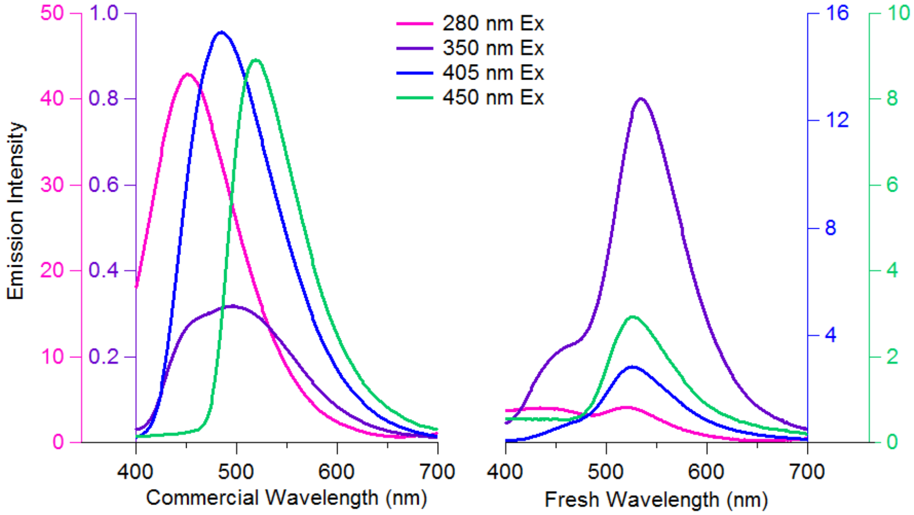


Figure 5.5. Spectral average differences between commercially purchased (left; N=29) and freshly collected (right; N=17) *Taraxacum officinale* pollen particles. The excitation/emission axes are colorized and conserved, per excitation, across each of the two types.

Figure 5.6 shows the differences in silver birch, *B. pendulam*, pollen commercially purchased from Allergon AB (Ängelholm, Sweden) and freshly collected from the Denver Botanic Gardens (Denver, USA). There are clear differences in the relative emission intensities of some of the *B. pendula* curves, but the fluorophore responses between each type seem conserved. It is important to note that this commercial set is reported by Allergon AB as not being defatted. Each of the two sets are dominated by fluorophore modes I, IV, V, and VI, in varying amounts. Other than relative intensity differences, however, there appears to be little difference between the fluorophore variation. In contrast to the previous set, *B. pendula* is an anemophilous species, and was explicitly reported as not defatted by the distributor.

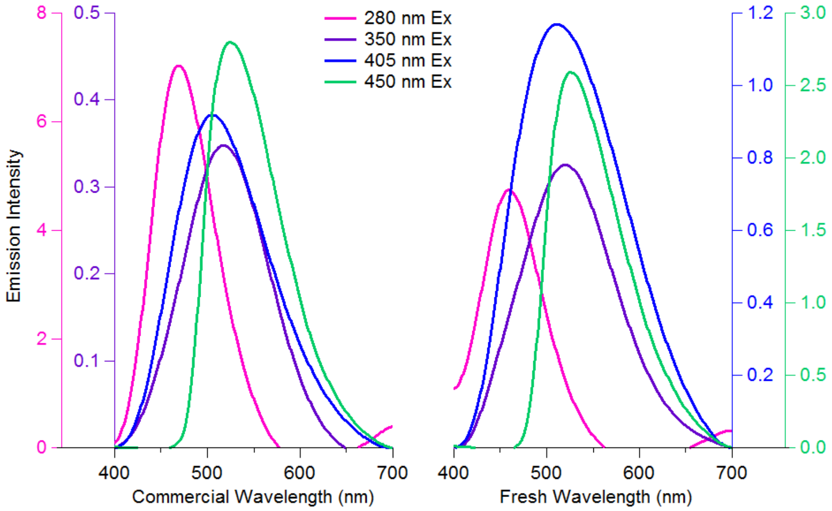


Figure 5.6. Spectral average differences between commercially purchased (left; N=24) and freshly collected (right; N=29) *Betula pendula* pollen particles. The excitation/emission axes are colorized and conserved, per excitation, across each of the two types.

The differences in these two sets exemplify that there are differences in both commercial and freshly collected sets. The degree to which these differences are affected by defatting, regional or local variation, or otherwise, is difficult to separate out. With these two samples, there is indication that, in some cases, defatting processes from commercial pollen development can affect the fluorescent properties to a larger degree than other differences due to region or soil quality.

5.5.2 Pollen Viability in Storage

Many previous pollen measurement studies were performed with commercially purchased pollen, and much this pollen was stored over a period in refrigerated conditions. Considering pollen morphology and viability can change very quickly (Schoper et al. 2010; Huang et al. 2004; Báez et al. 2002), even in refrigerated pollen samples, it is possible that other properties such as fluorescence may be affected as well.

Similarly, for the pollen inspected in chapter 4's clustering trials, the pollen was stored refrigerated for 1.5+ years. One species was collected in large amounts during the botanic gardens collections due to the prolific amounts of pollen it produced: *Pinus ponderosa*. It was not possible to examine identical particles in each sample, though particles from the same sample were analyzed in three time periods. This is shown in Figure 5.7, where the first analysis took place day after collection, the second was five months after collection, and third was seven months after the collection date. Clear differences are seen in the emission characteristics of each, though the seventh month behaves particularly strange in this regard. It is worth noting that each excitation sees a large increase in month 5, and subsequently lowers again in month 7, though it is not currently known why this is the case aside from the low sample size. Despite the small sample size, this may indicate the importance of taking sample measurements as quickly as possible after collection to get a representative sample, though a more comprehensive analysis of this will need to be done to understand the full impact of storage time.

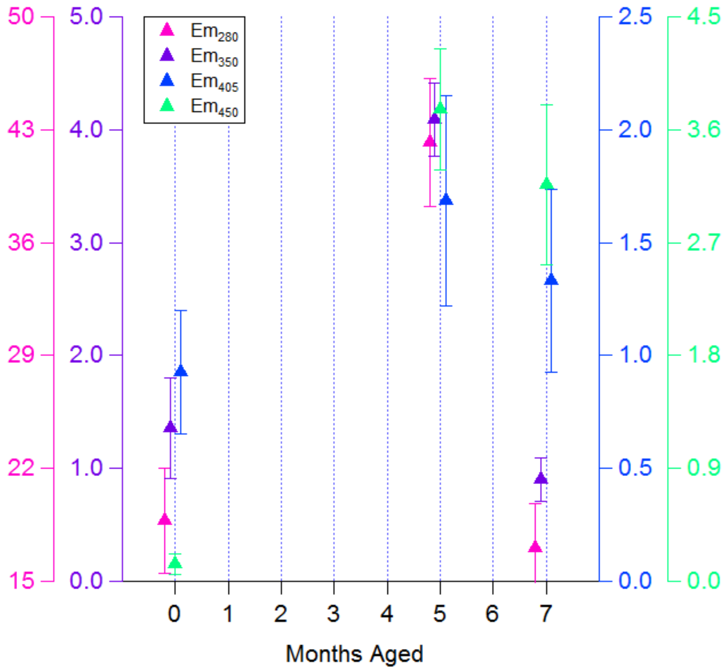


Figure 5.7. Spectral average differences for *Pinus ponderosa* over the course of seven months of refrigerated storage. The average for each set of monthly measurements are shown for each individual excitation on its own axis, with the maximum peak plotted for each emission type.

5.5.3 Collections across the Growing Season

Some plants also pollinate for a lengthy amount of time during the year. Very few species do this, so the options for this type of analysis were limited. Red-seeded dandelion pollen, *T. erythrospermum*, tends to have an extended pollination season, showing up from the early spring through the summer in Denver. Pollen for this species was collected in April, May, July, and September and subsequently analyzed on the sensor. These collections all took place on the south side of the University of Denver campus, and were analyzed within one to two days of collection and stored in refrigerated conditions. The max peak emission intensity for each set of collection averages are shown in figure 5.8 for each collection. Changes were noticed throughout the collection season, though the relatively large distributions on the emission intensities indicate that these particles can vary quite extensively. In particular, the 280 nm emission in July and September are particularly large, as well as the 450 emission for April and September, and the 405 emission for September. Aside from the final emission peak from 280 nm excitation in September, the emission peaks along each month were consistent.

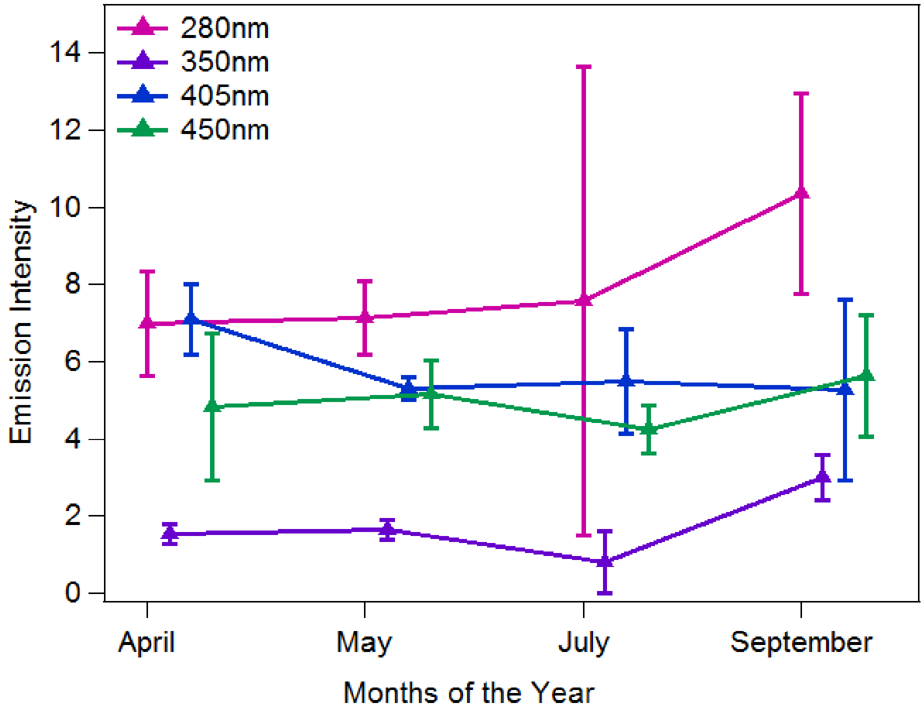


Figure 5.8. Emission peak max intensity of *Taraxacum erythrospermum* for each individual excitation-emission pair over four separate collections on differing months.

5.6 Ambient Data Predictions

The developed RF can be used as a framework to predict the similarity of collected ambient particles. Ambient particles were collected from the Denver Botanic Gardens on the same day of collection as both *B. pendula* and *P. tremuloides* in early 2019. Ambient collections were taken as described previously, with a pump and impactor, onto optical slides and analyzed by the sensor. A preliminary number of 9 particles were examined in this sample to show a proof-of-concept for prediction of new particles to the developed RF model. This is shown in Figure 5.9, which has three individual particle micrographs, along with their size and spectral information. Figure 5.9a shows one of the standards collected directly from a *B. pendula* tree, and shows the scattering image and measured parameters for this particle. Figure 5.9b represents a particle that was classified as a Carex pollen particle, which the standards for were collected on May 21st, 2018. Figure 5.9c represents a pollen particle classified into the *B. pendula* cluster from collection earlier that day.

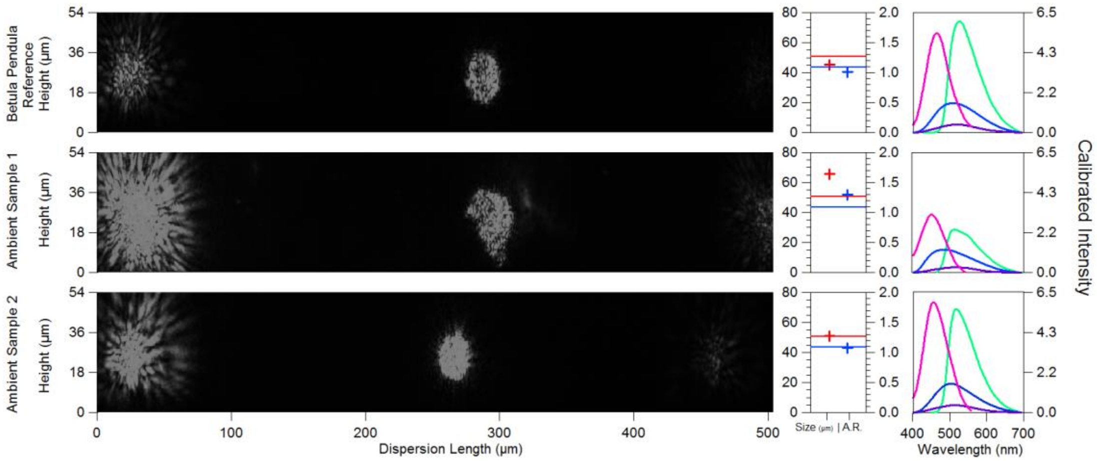


Figure 5.9. Several particles collected on the same day shown as their calibration scattering image, the size and aspect ratio for this particle (with the averages for *Betula pendula* listed as lines), and the emission spectra for all four excitations. (Top row) A standard *Betula pendula* particle collected directly from a tree at the Botanic gardens. (Middle row) One ambient particle that was classified as a *Carex* particle, and (Bottom row) a second ambient particle that was classified as a *Betula pendula* particle.

Comparisons of Figure 5.9a to 5.9c exemplify the similarities in these two types of particles, despite the bottom being collected from several hundred feet away from any *B.*

pendula trees. It isn't possible to determine, without outside confirmation, that these particles are from the same species. Still, the ambient particle shown here has identical size to the *B. pendula* average, and the spectral characteristics exhibited this particle to the species average is very similar. As such, classification of "*B. pendula*-like particle" may be a more apt label. Similarly, the middle row was classified as a *Carex*-like particle, which could be pollinating during this specific time, and exhibits a similar size/shape and spectral characteristics to previously measured *Carex* particles. The other 7 particles chosen for analysis on this slide were all classified as *B. pendula*-like particles as well.

5.7. Conclusion

Many previous UV-LIF sampling campaigns that have examined bioaerosols, such as pollen, have utilized commercially purchased or lab-grown samples. This has helped get some idea of how different bioaerosols may look using the suite of UV-LIF instrumentation, though it is hard to compare these measurements to real-world samples. Some studies have looked at the differences in freshly-collected pollen (Hernandez et al. 2016; O'Connor et al. 2011), though this has been a relatively recent effort with respect to UV-LIF instrumentation. In this chapter, we examined collections of fresh pollen made at the Denver Botanic Gardens and University of Denver campus over the course of the 2018 and early 2019 growing seasons. Species-level comparisons were made with two types of pollen, and the fluorescence emission characteristics showed that commercial pollen may not make the best comparisons with ambient sampling of pollen. Many anemophilous pollen types were collected, including pollen from nine grasses. Interestingly, this freshly collected grass pollen did not show signs of chlorophyll *a* fluorescence, which had previously been observed in several grass pollen samples from other studies (D. O’Connor et al. 2014a). Collections across multiple months were made for certain species to analyze fluorescence emission changes, as well as the effect that refrigeration storage has on the same parameters.

In total, 34 species were collected and analyzed via the random forest classification techniques developed in chapter 4. These techniques were used to develop a classification model for the entire pollination season, as well as seasonal subsets to account for potential pollination overlap of multiple species. This set was classified at the species level, the plant type as listed by the USDA, as well as the level of allergenicity for these pollen types. These developed seasonal models were then used to predict ambiently collected pollen particles at the Denver Botanic Gardens, and subsequently predict eight individual pollen grains to a species that was known to be pollinating hundreds of feet away at the time.

The data presented in this chapter gives the framework for predicting ambiently collected pollen to random forest models developed from pollen standards. Separatory power for this has been shown to be very strong, showing classification accuracy to be over 90% in some cases. Classification by certain groupings, such as USDA plant type, has been shown to increase this accuracy to as high as 98%. Classification of individual particles collected ambiently from hundreds of feet away has been shown to be possible.

Still, larger numbers of pollen types, both number per species and differing species, will need to be collected and analyzed on this system. To do this, advancements in the sampling techniques are needed. A real-time or semi-real-time system with a rolling tape and pump for sampling will need to be developed and implemented to sample large numbers of pollen particles for creating a catalogue to develop the random forest model with. Automated analysis techniques will also need to be developed for both image analysis and subsequent random forest classification and model prediction. This chapter, and those prior, describe the physical sensor and computational framework for the ability to classify and report pollen particle concentrations in the atmosphere.

Chapter 6: Conclusions

6.1 Summary of Conclusions

This thesis presents the development of a newsingle-particle fluorescence spectrometer for bioaerosol analysis, applied here to pollen, as well as various statistical and computational techniques to classify the data coming out of the instrument. The instrument operates similar to a slitless spectrometer, with the addition of excitation sources, a transmission grating, and a CCD camera detector to obtain fluorescence spectra of super-micron particles in a relatively inexpensive way. Many particles can be excited and detected simultaneously, allowing for multiple spectra to be obtained in an individual sample. This ultimately will allow for both a quick diagnostic (fluorescence or not) for particles, as well as in-depth probing of fluorophore response to an excitation source.

Development and characterization of the instrument happened iteritvely and simultaneously. Over the course of several years, new components were added (three excitation modes and an improved CCD camera) to improve detection and analysis capabilities. These improvements expanded the potential of applying fluorescence analysis from only carotenoid compounds and chloyphyll to add the ability to detect fluorophores from proteins and other phenolic structures as well. The new optical sources allow the instrument to probe similar information as a desktop fluorometer, though at a drastically reduced cost. This includes a range of excitation from 280-450 nm, with a reliable emission detection range of 400-700 nm. A size measurement scheme to measure the raw shape of the particles at their fluorescent angle was also implemented. The ability to normalize fluorescence by optical defects and excitation power density was also developed to ensure fair comparison between particles.

Clustering and classification techniques were tested with the data obtained through the instrument as well, resulting in the implementation of random forest classification. Four techniques in total (*k*-means; hierarchical agglomerative clustering; gradient boosting classification; random forest classification) were all tested against a data set of commercial pollen, with the latter two performing the best. Four-fold cross validation tests were performed using the two classification techniques, to nearly identical results. For computation efficiency, random forest was continued with further. Steps were taken to improve the data utilization, such as artificially weighting the size parameters, as well as to examine how important each source was in the overall developed training model. This was coupled with examining how reducing the number of sources changed the model accuracy, showing that the 280 and 350 nm sources were of particularly high important in the models, as well as the size.

Collection of fresh pollen samples was performed over the course of a year at the Denver Botanic Gardens and on the University of Denver campus. This resulted in the collection of 34 species and analysis of 932 total pollen particles. Classification of these particles at the species level showed 90% accuracy, and subsequent classification at different levels (allergenicity, plant type, etc) as well as limiting by possible overlapping of pollination times (i.e. plants that pollinate in a given month are unlikelyto also pollinate during the same month as plants that pollinate several months later) showed marginal increases (4-8%) in classification. Initial classifications of ambiently collected species were also tested, and able to classify as a type of particle that was known to be pollinating across the Botanic Gardens on the same day.

6.2 State of Instrumental Application to Pollen Analysis

This thesis has shown the development of both the single-particle fluorescence instrument as well as the techniques being used to analyze data from the instrument. As such, is an inexpensive (e.g. <$6000) pollen detection system that is able to classify pollen to the species level with high accuracy, and with a collection set of 34 species that accuracy was 90%. Classification by allergenicity and reducing the overlap of unlikely co-pollinating species increases the accuracy by a few percentage points. The random forest classifications were performed with a relatively low sample number per species (~25/species) all from a very similar location (Denver area, mostly the Botanic Gardens), and mostly from single plants or plants in the same area even within the Botanic Gardens. It is possible that different environments and growing conditions may affect spectral characteristics in a myriad of ways. However, one cross-species example between commercially grown in Europe and freshly collected in the United States, Silver Birch, was shown to have extremely similar fluorescence characteristics. This bodes well for potential intraspecies differences in fluorescence for pollen in different areas and environments.

Compared to traditional techniques, this instrument has the potential to eliminate the need for a technician, or palynologist for traditional methods, as the process from collection to analysis can be automated in some way. Most of the individual components, aside from collection itself, have been individually automated in some way. For example, image analysis has been taken from individual analysis and manual calibration to a fully automated process through the Igor programming language that outputs finalized spectra for several sets of images simultaneously. Similarly, the ability to operate the fluorescence collection has been partially automated. Assuming a collection system that allows for semi-continuous sampling, it may be possible to sample fluorescence properties of aerosol particles in 5-10-minute batches, drastically increasing resolution from traditional techniques which sample over the course of a week. Since all of the spectral statistics are done via algorithmic classification, not visual comparison, the process of classifying newly collected particles to groups can be done quickly and reliably during a subsequent sample collection.

The instrument can accurately detect fluorescence from pollen and classify it to high (>90%) accuracy by allergenicity, plant-type, and even to the species level. While there are commercially available instruments to sample atmospheric pollen such as the KH-3000-1, though the price (frequently >$100k) will prevent widespread usage. Though this instrument will never be able to detect particles at the same overall scale of particle numbers per instrument, not being a real-time method, the analysis of several particle simultaneously as well as the ability to deploy many more instruments per cost (~15X). Considering the high level of separation achieved with classification by allergenicity

(>92%), deploying many of these instruments around a single city may enable close to real-time reporting of allergenic pollen concentrations in the air.

There are smaller, commercial platforms that allow for pollen detection, though they rely on visual microscopy techniques similar to traditional classification methods. The PollenSense instrument identifies pollen based on images/morphology, as does the instrument developed by Wu et al. that has not been commercialized yet (Wu et al. 2018). Instruments that differentiate pollen species based on morphology alone are not generally capable of detecting subtle differences in pollen groups that may have very similar morphology, such as some grass pollens (Mander et al. 2014). The instrument was used to differente between several pollen of a single species (especially those in the *Pinus* genus, a tree pollen, for example), though it remains to be seen if there are significant differences in other types (grass, forb, ect). If the collection of fluorescence information can further improve classification beyond what is capable by visual microscopy, then this technique could be an important complementary tool to supplement existing detection techniques.

6.3 Current Limitations and Future Steps

In its current state, the instrument is a desktop-affixed instrument that needs a technician to operate. Many of the processes including the post-spectral collection analysis, calibration, and classification analysis have been individually automated, though these steps have not been bridged together completely. The spectral collection itself can be fully automated, though not in the present configuration. One of the most time-consuming sections of the overall analysis is finding individual particles in sparse samples, or finding particles that are representative of the sample, and properly getting it into the CCD viewable area. Currently, these images are positioned by the instrument operator. Since particle locations on the sample slide may be variable, resulting in areas of more, or less, pollen density, a method to steadily roll across the slide surface and systematically image it will need to be developed. The stage could be attached to a small motor that consistently moves slide over after a set of images are taken, allowing for some overlap (i.e. it moves the slide only far enough that 80% of the scene is now new, with 20% from the last set) to compensate for particles on the edge of the viewing area. Associated with finding particle locations is figuring out how fluorescent a particle, or particle type, is. Many similar sized pollen particles have shown extremely variable fluorescence signals that may require different exposure times to collect adequate spectra (i.e. spectra that isn't too dark or saturated). Some progress has been made to alleviate this, with image recognition python code to automatically detect thresholds associated with particles in the image, though this is still a preliminary process. In the future, operating the instrument in two modes, low and high exposure, may allow for the detection of multiple types of particles in a single sample. Dual-exposure operation is likely key, since atmospheric samples will contain more diversity than pollen alone, and the ability to differentiate between background contaminants like dust, spores, or pollen will be extremely important.

No collection mechanism exists on the platform currently. Since it is attached to a breadboard, the instrument could be transported outdoors, though the absence of a collection mechanism integrated into the system (i.e. without the need for human interaction) makes this a pointless effort. All particles either need to be collected from the source plant, collected via deposition from atmospheric particles onto the slide (i.e. leaving it outside for a period of time) or by tapping a sample onto the slide, or collected with an impactor onto a slide to be put into the system later. Future collection may involve scrapping the optical slides in favor of a nonfluorescent rolling tape, similar to the resourse effective bioidentification system from Battelle (Doughty and Hill 2017), that allows for continuous, semi-real-time measurements. To do this, sampling time and flow rate will need to be optimized for the larger sized pollen particles. Collection also needs to currently be monitored carefully, as both deposition and impaction can result in samples that are either too sparse or too dense. Samples that are too dense (i.e. streaks overlap) cannot currently be used, as no method to separate them has been implemented. To combat this, multi-peak fitting or positive matrix factorization will need to be implemented to deconvolute spectral signals from overlapping particles.

Though the RF classification showed high classification accuracy, and a preliminary attempt to classifying ambiently collected particles was successful, this technique requires a library of particles to be utilized to develop the RF model. Needing a library of standard particles presents a number of problems. The first problem is that a large library of pollen is needed for a model to be effective in classifying ambiently collected pollen. Pollen fluorescence emission can vary significantly between individual species, as well as within a single species, indicating that both the types of pollen and number of pollens per species needs to be increased in these models. Since the models discussed here only encompass 34 pollen species in total, any predictions of ambient particles can be reported by similarity, not by actual identity. Without co-location of traditional pollen sampling techniques, there is no way to properly test the efficacy of these classification techniques.

6.4 Overall Analysis of Pollen Detection

The instrument and application presented here provide a path forward for pollen detection and classification for relatively inexpensive cost. Preliminary work shows that the collections of pollen analyzed can be accurately separated (>90%) using random forest classification. Narrowing these classifications to pollen by type or allergenicity can increase accuracy even further. High classification accuracy for the pollen types shown here have better accuracy than traditional methods (Mander et al. 2014), and the focus on pollen that may be found in the front range lend credence to the technique's ability to classify pollen in an individual area. Collections and classification of ambient pollen was also assigned by the model as a species known to be pollinating on the same day at a distance away. If multiple instrument units are deployed simultaneously as a small-scale network, it may bridge the gap both spatially (in between current sites) and temporally (since current sites do not report directly) towards improved allergen forecasting.

UV-LIF detection of pollen has been a useful tool to supplement traditional techniques, though there have been many challenges since most of the commercial instrumentation has been applied mainly to smaller biological aerosols (Savage and Huffman 2018; Hernandez et al. 2016; O'Connor et al. 2011; Hill et al. 1999), making it hard to analyze pollen effectively. Newer technologies like the Pollen-Sense[d], and other image-recognition instrumentation (Wu et al. 2018), and our instrument that have focused on pollen detection and classification have begun to bridge the gaps that commercial UV-

LIF rarely could. Considering these recent advancements, allergy monitoring, and prediction is poised to experience a boom where the data input can be reported in semi-real time (on the order of hours), in contrast to the current approach. For this to happen, research groups and commercial entities will have to collaborate on the creation of databases for these types of instruments, as local and regional differences in pollinating species will be important for accurate reporting.

This thesis discusses the instrumentation and its application to pollen alone. Both the instrument itself, as well as the subsequent techniques for classification, in principle can be applied to other bioaerosols or fluorescent particle types. For the instrument itself, both the sampling and collection techniques can be modified to focus on smaller particle types like fungal spores. Filtering out larger particles from the sampling process and increasing the magnification of the optics, or increasing the exposure time of the camera, will allow fungal spores, which typically range from 1-10 am, to be analyzed instead of the relatively larger pollen particles. The classification techniques described here can be applied to any instrumentation and have been applied to WIBS and SIBS data prior but are not described here. Pairing this instrument with real-time commercial instrumentation like the WIBS may give insight into nuances between these types of techniques outside of the obvious spectral- and time-resolution differences.

Though the instrument presented here is able to acquire highly-resolved fluorescence spectral signals at a fairly low cost, there are several steps that need to take place prior to usage in pollen reporting. The first step is to integrate a collection mechanism that enables semi-continuous sampling of pollen. Once that is incorporated, it will be important to connect the system to a Raspberry Pi, or other inexpensive computational platform, and work towards automation of the system so collection, detection, analysis, and classification can be performed autonomously. These goals are achievable from an engineering and computational standpoint and will require input of time and effort.

The largest unknown is that of the detection and classification system for ambient particle collections, and development of models that are able to encompasse enough unknowns to be viable in reporting pollen forecasts. As a part of the work discussd here, 34 species of pollen were collected directly from trees in the Denver Bonatic Gardens and are treated here as ‘plant standards’. There was discrimination while searching for particles on these slides, as to which to include, in order to limit interferences that may present a more realistic scenario. However, limiting these variables was done to provide a proof of concept that the technique could be useful at all. In the future, much larger models will need to be developed with many species of pollinating plants and will also need to include types of particles that may be interferences. Particles that are predicted by the RF models can be reported by their identity (i.e. “this particle is *B. pendula”*) or by their percentage similarity to each model’s cluster. This similarity percentage can be used to develop thresholding strategies for predictions. For example, particles that are reported as a certain species, but only seem loosely related based on their percentage of similarity, may help throw out predictions that are likely erroneous in some way. These types of strategies will be the crux of the overall technique’s application.

If the technique presented here is able to detect and classify pollen to the species level, it will represent a drastic increase in the ability to report current pollen levels to the public. If it is able to only detect allergenicity of the pollen collected, this advancement is still very important. Seasonal pollen allergies account for a significant percentage of daily health issues in the world and having a more accurate model of pollen allergens in the atmosphere, both spatially and temporally, will help the public more effectively prepare in their day to day life. The inexpensive nature of this technique may allow widespread coverage of pollen allergen reporting than previously possible.

References

Ackerman, J.D. (2000). Abiotic pollen and pollination: Ecological, functional, and evolutionary perspectives. *Plant Syst. Evol.,* 167–185.

Allen, G.P., Hodgson, R.M., Marsland, S.R., and Flenley, J.R. (2008). Machine vision for automated optical recognition and classification of pollen grains or other singulated microscopic objects., *in 15th International Conference on Mechatronics and Machine Vision in Practice, M2VIP'08,* pp. 221–226.

Ariya, P.A., Sun, J., Eltouny, N.A., Hudson, E.D., Hayes, C.T., and Kos, G. (2009). Physical and chemical characterization of bioaerosols - Implications for nucleation processes. *Int. Rev. Phys. Chem.,* 28(1):1–32.

Aronne, G., Cavuoto, D., and Eduardo, P. (2001). Classification and counting of fluorescent pollen using an image analysis system. *Biotech. Histochem.,* 76(1):35–40.

Asero, R., Wopfner, N., Gruber, P., Gadermaier, G., and Ferreira, F. (2006). Artemisia and Ambrosia hypersensitivity: Co-sensitization or co-recognition? *Clin. Exp. Allergy,* 32(7):667–670.

Atkins, P. and De Paula, J. (1989). Atkin's Physical Chemistry, *Journal of Chemical Information and Modeling*.

Axelrod, D., Koppel, D.E., Schlessinger, J., Elson, E., and Webb, W.W. (1976). Mobility measurement by analysis of fluorescence photobleaching recovery kinetics. *Biophys. J.,* 16(9):1055–1069.

Aylor, D. (1975). Deposition of particles in a plant canopy. *J. Appl. Meteorol.,* 14(1):52–57.

Báez, P., Riveros, M., and Lehnebach, C. (2002). Viability and longevity of pollen of Nothofagus species in south Chile. *New Zeal. J. Bot.,* 40(4):671–678.

Balková, E. (2015). Oral allergy syndrome and pollen food allergy syndrome. *Lek. Obz.,* 26(2):78–88.

Bartlett, J. (2008). Bioaerosols Handbook, *Occupational and Environmental Medicine*.

Bennett, K.D. and Willis, K.J. (2002). Pollen, *Tracking Environmental Change Using Lake Sediments*.

Boyain-Goitia, A.R., Beddows, D.C.S., Griffiths, B.C., and Telle, H.H. (2003). Single-pollen analysis by laser-induced breakdown spectroscopy and Raman microscopy.

Appl. Opt., 42(30):6119–6132.

Bragg, L.H. (1969). Pollen Size Variation in Selected Grass Taxa. *Ecology,* 50(1):124–127.

Breiman, L. (2001). Random Forrests. *Mach. Learn.,* 45(1):5–32.

Buckland, M. and Gey, F. (1994). The relationship between Recall and Precision. *J. Am. Soc. Inf. Sci.,* 45(1):12–19.

Bunch, B.H. and Hellemans, A. (2004). The history of science and technology: a browser's guide to the great discoveries, inventions, and the people who made them from the dawn of time to today, *Choice Reviews Online*. Houghton Mifflin Harcourt.

Bünger, J., Antlauf-Lammers, M., Schulz, T., Westphal, G., Müller, M., Ruhnau, P., and Hallier, E. (2000). Health complaints and immunological markers of exposure to bioaerosols among biowaste collectors and compost workers. *Occup. Environ. Med.,* 57(7).

Bureau, A., Dupuis, J., Hayward, B., Falls, K., and Van Eerdewegh, P. (2003). Mapping complex traits using Random Forests. *BMC Genet.,* 4(1):S64.

Bureau, N.A. (2019). NAB Pollen Counts, *National Allergy Bureau*.

Buters, J., Antunes, C., Galveias, A., Bergmann, K., Thibaudon, M., Galán, C., Schmidt-Weber, C., and Oteros, J. (2018). Pollen and spore monitoring in the world. *Clin. Transl. Allergy,* 8(1):9.

Calvo, A., Baumgardner, D., Castro, A., Fernández-González, D., Vega-Maray, A., Valencia-Barrera, R., Oduber, F., Blanco-Alegre, C., and Faile, R. (2018). Daily behavior of urban Fluorescing Aerosol Particles in northwest Spain. *Atmos. Environ.,* (184):262–277.

Chen, C., Hendriks, E.A., Duin, R.P.W., Reiber, J.H.C., Hiemstra, P.S., De Weger, L.A., and Stoel, B.C. (2006). Feasibility study on automated recognition of allergenic pollen: Grass, birch and mugwort. *Aerobiologia (Bologna).,* 22(4):275–284.

Cheng, J., Liu, Y., Cheng, X., He, Y., and Yeung, E.S. (2010). Real time observation of chemical reactions of individual metal nanoparticles with high-throughput single molecule spectral microscopy. *Anal. Chem.,* 82(20):8744–8749.

Cohen, S.H., Yunginger, J.W., Rosenberg, N., and Fink, J.N. (1979). Acute allergic reaction after composite pollen ingestion. *J. Allergy Clin. Immunol.,* 64(4):270–274.

Crawford, I., Ruske, S., Topping, D., and Gallagher, M. (2015). Evaluation of hierarchical agglomerative cluster analysis methods for discrimination of primary biological aerosol. *Atmos. Meas. Tech.,* 8(11):4979–4991.

Crouzy, B., Stella, M., Konzelmann, T., Calpini, B., and Clot, B. (2016). All-optical automatic pollen identification: Towards an operational system. *Atmos. Environ.,* 140:202–212.

D'Amato, G., Cecchi, L., Bonin, S., Nunes, C., I., A., Behrendt, H., Liccardi, G., Popov, T., and Cauwenberge, P. Van (2007). Allergenic pollen and pollen allergy in Europe. *Allergy,* 62(9):976–990.

Daszykowski, M. and Walczak, B. (2010). Density-Based Clustering Methods., *in Comprehensive Chemometrics, .*

Dell'Anna, R. (2010). A critical presentation of innovative techniques for automated pollen identification in aerobiological monitoring networks., *in Pollen: Structure, Types, and Effects,* pp. 273–288.

Després, V., Huffman, J.A., Burrows, S., Hoose, C., Safatov, A., Buryak, G., Fröhlich-Nowoisky, J., Elbert, W., Andreae, M., Pöschl, U., and Jaenicke, R. (2012). Primary biological aerosol particles in the atmosphere: a review. *Chem. Phys. Meteorol.,* 64(1):15598.

Diethart, B., Sam, S., and Weber, M. (2007). Walls of allergenic pollen: Special reference to the endexine. *Grana,* 46(3):164–175.

Dietterich, T. (1995). Overfitting and undercomputing in machine learning. *ACM Comput. Surv.,* 27(3):326–327.

Dobrucki, J.W. and Kubitscheck, U. (2017). Fluorescence Microscopy., *in Fluorescence Microscopy: From Principles to Biological Applications: Second Edition, .*

Doughty, D. and Hill, S. (2017). Automated aerosol Raman spectrometer for semi-continuous sampling of atmospheric aerosol. *J. Quant. Spectrosc. Radiat. Transf.,* 188:103–117.

Douwes, J., Thorne, P., Pearce, N., and Heederik, D. (2003). Bioaerosol health effects and exposure assessment: Progress and prospects. *Ann. Occup. Hyg.,* 47(3):187–200.

Dua, K.L. and Shivpuri, D.N. (1962). Atmospheric pollen studies in Delhi area in 1958-1959. *J. Allergy,* 33(6):507–512.

Elangasinghe, M.A., Singhal, N., Dirks, K.N., Salmond, J.A., and Samarasinghe, S. (2014). Complex time series analysis of PM10 and PM2.5 for a coastal site using artificial neural network modelling and k-means clustering. *Atmos. Environ.*,.

Erdmann, N., Dell'Acqua, A., Cavalli, P., Grüning, C., Omenetto, N., Putaud, J.P., Raes, F., and Van Dingenen, R. (2005). Instrument characterization and first application of the single particle analysis and sizing system (SPASS) for atmospheric aerosols. *Aerosol Sci. Technol.,* 39(5):377–393.

Fennelly, M.J., Sewell, G., Prentice, M.B., O'Connor, D.J., and Sodeau, J.R. (2017). Review: The use of real-time fluorescence instrumentation to monitor ambient primary biological aerosol particles (PBAP). *Atmosphere (Basel).,* 9(1):1.

Forde, E., Gallagher, M., Foot, V., Sarda-Esteve, R., Crawford, I., Kaye, P., Stanley, W., and Topping, D. (2019). Characterisation and source identification of biofluorescent aerosol emissions over winter and summer periods in the United Kingdom. *Atmos. Chem. Phys.,* 19(3):1665–1684.

Franze, T., Weller, M.G., Niessner, R., and Pöschl, U. (2005). Protein nitration by polluted air. *Environ. Sci. Technol.,* 39(6):1673–1678.

Friedman, J.H. (2011). Greedy function machine: A gradient boosting machine. *Statistics (Ber).*,.

Friedman, J.H. (2001). Greedy function approximation: A gradient boosting machine. *Ann. Stat.,* 29(5):1189–1232.

Fröhlich-Nowoisky, J., Kampf, C., Weber, B., Huffman, J.A., Pöhlker, C., Andreae, M., Lang-Yona, N., Burrows, S., Gunthe, S., Elbert, W., Su, H., Hoor, P., Thines, E., Hoffmann, T., Després, V., and Pöschl, U. (2016). Bioaerosols in the Earth system: Climate, health, and ecosystem interactions. *Atmos. Res.*,.

Gabey, A.M., Gallagher, M.W., Whitehead, J., Dorsey, J.R., Kaye, P.H., and Stanley, W.R. (2010). Measurements and comparison of primary biological aerosol above and below a tropical forest canopy using a dual channel fluorescence spectrometer. *Atmos. Chem. Phys.,* 10(10):4453–4466.

Galán, C., García-Mozo, H., Cariñanos, P., Alcázar, P., and Domínguez-Vilches, E. (2001). The role of temperature in the onset of the Olea europaea L. pollen season in southwestern Spain. *Int. J. Biometeorol.,* 45(1):8–12.

Geburek, T., Hiess, K., Litschauer, R., and Milasowszky, N. (2012). Temporal pollen pattern in temperate trees: Expedience or fate? *Oikos,* 121(10):1603–1612.

Górny, R.L., Dutkiewicz, J., and Krysińska-Traczyk, E. (1999). Size distribution of bacterial and fungal bioaerosols in indoor air. *Ann. Agric. Environ. Med.,* 6:105–113.

Górny, R.L., Reponen, T., Willeke, K., Schmechel, D., Robine, E., Boissier, M., and Grinshpun, S.A. (2002). Fungal fragments as indoor air biocontaminants. *Appl. Environ. Microbiol.*, 68(7):3522–3531.

Green, B.J., Tovey, E.R., Sercombe, J.K., Blachere, F.M., Beezhold, D.H., and Schmechel, D. (2006). Airborne fungal fragments and allergenicity. *Med. Mycol.,* 44(Supplement 1):S245–S255.

Grote, M., Valenta, R., and Reichelt, R. (2003). Abortive pollen germination: A mechanism of allergen release in birch, alder, and hazel revealed by immunogold electron microscopy. *J. Allergy Clin. Immunol.,* 111(5):1017–1023.

Gruijthuijsen, Y.K., Grieshuber, I., Stöcklinger, A., Tischlera, U., Fehrenbach, T., Weller, M.G., Vogel, L., Vieths, S., Pöschl, U., and Duschl, A. (2006). Nitration enhances the allergenic potential of proteins. *Int. Arch. Allergy Immunol.,* 141(3):265–275.

Hairston, P.P., Ho, J., and Quant, F.R. (1997). Design of an instrument for real-time detection of bioaerosols using simultaneous measurement of particle aerodynamic size and intrinsic fluorescence. *J. Aerosol Sci.,* 28(3):471–482.

Han, H., Guo, X., and Yu, H. (2017). Variable selection using Mean Decrease Accuracy and Mean Decrease Gini based on Random Forest., *in Proceedings of the IEEE International Conference on Software Engineering and Service Sciences, ICSESS,* pp. 219–224.

Hartigan, J.A. and Wong, M.A. (2006). Algorithm AS 136: A K-Means Clustering Algorithm. *Appl. Stat.,* 28(1):100–108.

Healy, D.A., O'Connor, D.J., Burke, A.M., and Sodeau, J.R. (2012). A laboratory assessment of the Waveband Integrated Bioaerosol Sensor (WIBS-4) using individual samples of pollen and fungal spore material. *Atmos. Environ.,* 60:534–534.

Hernandez, M., Perring, A., McCabe, K., Kok, G., Granger, G., and Baumgardner, D. (2016). Chamber catalogues of optical and fluorescent signatures distinguish bioaerosol classes. *Atmos. Meas. Tech.,* 9(7).

Hesse, M. (1984). Pollenkitt is lacking in Gnetatae:Ephedra and Welwitschia; further proof for its restriction to the angiosperms. *Plant Syst. Evol.,* 144(1):9–16.

Hesse, M. (1981). The fine structure of the exine in relation to the stickiness of angiosperm pollen. *Rev. Palaeobot. Palynol.,* 35(1):81–92.

Hill, SC, Pinnick, R., Niles, S., and Pan, Y. (1999). Real-time measurement of fluorescence spectra from single airborne biological particles. *F. Anal.*,.

Hill, S, Pinnick, R., Niles, S., Pan, Y.L., Holler, S., Chang, R., Bottinger, J., Chen, B., Orr, C.S., and Feather, G. (1999). Realtime Measurement of Fluorescence Spectra from Single Airborne Biological Particles. *F. Anal. Chem. Technol.*, 3(4–5):221–239.

Hill, S., Williamson, C., Doughty, D., Pan, Y. Le, Santarpia, J., and Hill, H. (2015). Size-dependent fluorescence of bioaerosols: Mathematical model using fluorescing and absorbing molecules in bacteria. *J. Quant. Spectrosc. Radiat. Transf.*, 157:54–70.

Hill, S.C., Mayo, M.W., and Chang, R.K. (2009). Fluorescence of Bacteria , Pollens , and Naturally Occurring Airborne Particles : Excitation / Emission Spectra. *Army Res. Lab. Appl. Phys.*, (No. ARL-TR-4722).

Hill, S.C., Pinnick, R.G., Niles, S., Fell, N.F., Pan, Y.-L., Bottiger, J., Bronk, B. V., Holler, S., and Chang, R.K. (2001). Fluorescence from airborne microparticles: dependence on size, concentration of fluorophores, and illumination intensity: erratum. *Appl. Opt.*, 40:3005–3013.

Hjelmroos, M. (1991). Evidence of long-distance transport of betula pollen. *Grana*, 30(1):215–228.

Hothorn, T., Lausen, B., Benner, A., and Radespiel-Tröger, M. (2004). Bagging survival trees. *Stat. Med.*, 23(1):77–91.

Hothorn, T., Zeileis, A., Cheng, E., and Ong, S. (2015). partykit: A modular toolkit for recursive partytioning in R. *J. Mach. Learn. Res.*, 16(1):3905–3909.

Hryhorczuk, D., Curtis, L., Scheff, P., Chung, J., Rizzo, M., Lewis, C., Keys, N., and Moomey, M. (2001). Bioaerosol emissions from a suburban yard waste composting facility. *Ann. Agric. Environ. Med.*, 8(2):177–185.

Huang, H.C., Pan, Y.-L., Hill, S.C., and Pinnick., R.G. (2011). Fluorescence-Based Classification with Selective Collection and Identification of Individual Airborne Bioaerosol Particles., *in Optical Processes In Microparticles And Nanostructures: A Festschrift Dedicated to Richard Kounai Chang on His Retirement from Yale University.*, pp. 153–167.

Huang, Z., Zhu, J., Mu, X., and Lin, J. (2004). Pollen dispersion, pollen viability and pistil receptivity in Leymus chinensis. *Ann. Bot.*, 93(3):295–301.

Huffman, D. and Huffman, J.A. (2019). A Wavelength Dispersive Microscope Spectrofluorometer for Characterizing Multiple Particles Simultaneously. US Patent US20160320306A1, filed January 8, 2014, and issues April 23, 2019.

Huffman, D., Swanson, B., and Huffman, J.A. (2016). A wavelength-dispersive instrument for characterizing fluorescence and scattering spectra of individual aerosol particles on a substrate. *Atmos. Meas. Tech.*, 9(8):3987–3998.

Huffman, J., Perring, A., Savage, N., Clot, B., Crouzy, B., Tummon, F., Shoshanim, O., Damit, B., Schneider, J., Sivaprakasam, V., Zawadowicz, M., Crawford, I., Gallagher, M., Topping, D., Doughty, D., Hill, S., and Pan, Y. Le (2019). Real-time sensing of bioaerosols: Review and current perspectives. *Aerosol Sci. Technol.*, 1–56.

Huffman, J.A. and Santarpia, J. (2017). Microbiology of Aerosols: Online Techniques for Quantification and Characterization of Biological Aerosols.

IMS Health Incorporated (2019). Pollen Library.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). An Introduction to Statistical Learning. New York: springer.

Jones, M.D. and Newell, L.C. (1948). Size, Variability, and Identification of Grass Pollen1. *J. Am. Soc. Agron.*,.

Karle, A.C., Oostingh, G.J., Mutschlechner, S., Ferreira, F., Lackner, P., Bohle, B., Fischer, G.F., Vogt, A.B., and Duschl, A. (2012). Nitration of the pollen allergen bet v 1.0101 enhances the presentation of bet v 1-derived peptides by HLA-DR on human dendritic cells. *PLoS One*, 7(2):e31483.

Kasprzyk, I. (2004). Airborne pollen of entomophilous plants and spores of pteridophytes in Rzeszów and its environs (SE Poland). *Aerobiologia (Bologna).*, 20(4):217–222.

Katelaris, C.H. (2010). Food allergy and oral allergy or pollen-food syndrome. *Curr. Opin. Allergy Clin. Immunol.*, 10(3):246–251.

Kawashima, S., Clot, B., Fujita, T., Takahashi, Y., and Nakamura, K. (2007). An algorithm and a device for counting airborne pollen automatically using laser optics. *Atmos. Environ.*, 41(36):7987–7993.

Kawashima, S., Thibaudon, M., Matsuda, S., Fujita, T., Lemonis, N., Clot, B., and Oliver, G. (2017). Automated pollen monitoring system using laser optics for observing seasonal changes in the concentration of total airborne pollen. *Aerobiologia (Bologna).*, 33(3):351–362.

Kaye, P., Stanley, W.R., Hirst, E., Foot, E. V, Baxter, K.L., and Barrington, S.J. (2005). Single particle multichannel bio-aerosol fluorescence sensor. *Opt. Express*, 13(10):3583–3593.

Khatun, S. and Flowers, T.J. (2007). The estimation of pollen viability in rice. *J. Exp. Bot.,* 46(1):151–154.

Kiselev, D., Bonacina, L., and Wolf, J.-P. (2011). Individual bioaerosol particle discrimination by multi-photon excited fluorescence. *Opt. Express,* 19(24):24516–24521.

Kiselev, D., Bonacina, L., and Wolf, J.P. (2013). A flash-lamp based device for fluorescence detection and identification of individual pollen grains. *Rev. Sci. Instrum.,* 84(3):033302.

Knobelspiesse, K.D., Pietras, C., Fargion, G.S., Wang, M., Frouin, R., Miller, M.A., Subramaniam, A., and Balch, W.M. (2004). Maritime aerosol optical thickness measured by handheld sun photometers. *Remote Sens. Environ.,*.

Könemann, T., Ditas, F., Walter, D., Brooks, J., Rodriguez-Caballero, E., Charlotte M. Dewald, M.D., Drewnick, F., Edtbauer, A., Fachinger, F., Harder, H., Klingmüller, K., Lammel, G., Paulsen, H., Weber, B., Wietzoreck, M., Williams, J., Yordanova, P., Borrmann, S., Lelieveld, J., Meinrat, O.A., Huffman, J.A., Pöschl, U., and Pöhlker, C. (2019a). Online analysis of single aerosol particle fluorescence spectra during the AQABA research cruise around the Arabian Peninsula. *Atmos. Chem. Phys.,* In Review.

Könemann, T., Savage, N., Huffman, J.A., and Pöhlker, C. (2017). Systematic characterization of fluorescence properties of polystyrene latex spheres using off- and on-line methods.No Title. (Technical Note).

Könemann, T., Savage, N., Klimach, T., Walter, D., Fröhlich-Nowoisky, J., Su, H., Pöschl, U., Huffman, J.A., and Pöhlker, C. (2019b). Spectral Intensity Bioaerosol Sensor (SIBS): An instrument for spectrally resolved fluorescence detection of single particles in real time. *Atmos. Meas. Tech.,* 12(2):1337–1363.

Kümmel, M., Walsh, J.R., Pirzkal, N., Kuntschner, H., and Pasquali, A. (2009). The Slitless Spectroscopy Data Extraction Software aXe. *Publ. Astron. Soc. Pacific,* 121(875):59.

Kuparinen, A. (2006). Mechanistic models for wind dispersal. *Trends Plant Sci.,* 11(6):296–301.

Kuparinen, A., Katul, G., Nathan, R., and Schurr, F.M. (2009). Increases in air temperature can promote wind-driven dispersal and spread of plants. *Proc. R. Soc. B Biol. Sci.,* 276(1670):3081–3087.

Lakowicz, J.R. (2006). Principles of fluorescence spectroscopy, *Principles of Fluorescence Spectroscopy*. Springer Science & Business Media.

Lee, M., Yaglidere, O., and Ozcan, A. (2011). Field-portable reflection and transmission microscopy based on lensless holography. *Biomed. Opt. Express,* 2(9):2721–2730.

Léonard, R., Wopfner, N., Pabst, M., Stadlmann, J., Petersen, B.O., Duus, J., Himly, M., Radauer, C., Gadermaier, G., Razzazi-Fazeli, E., Ferreira, F., and Altmann, F. (2010). A new allergen from ragweed (Ambrosia artemisiifolia) with homology to art v 1 from mugwort. *J. Biol. Chem.,* 285(35):27192–27200.

Leuschner, R.M. (1993). Pollen. *Experientia,* 49(11):931–942.

Liu, Y., Li, Z., Xiong, H., Gao, X., and Wu, J. (2010). Understanding of internal clustering validation measures.*, in Proceedings - IEEE International Conference on Data Mining, ICDM,* pp. 911–916.

Mackiewicz, B. (1998). Study on exposure of pig farm workers to bioaerosols, immunologic reactivity and health effects. *Ann. Agric. Environ. Med.,* (5):169–176.

MacQueen, J.B. (1967). Kmeans Some Methods for classification and Analysis of Multivariate Observations. *5th Berkeley Symp. Math. Stat. Probab. 1967,* 1(233):281–297.

Madonna, A.J., Voorhees, K.J., and Hadfield., and T.L. (2001). Rapid detection of taxonomically important fatty acid methyl ester and steroid biomarkers using in situ thermal hydrolysis/methylation mass spectrometry (THM-MS): implications for bioaerosol detection. *J. Anal. Appl. Pyrolysis,* 61(1–2):65–89.

Mander, L., Baker, S.J., Belcher, C.M., Haselhorst, D.S., Rodriguez, J., Thorn, J.L., Tiwari, S., Urrego, D.H., Wesseln, C.J., and Punyasena, S.W. (2014). Accuracy and Consistency of Grass Pollen Identification by Human Analysts Using Electron Micrographs of Surface Ornamentation. *Appl. Plant Sci.,* 2(8):1400031.

Miki, K., Kawashima, S., Clot, B., and Nakamura, K. (2019). Comparative efficiency of airborne pollen concentration evaluation in two pollen sampler designs related to impaction and changes in internal wind speed. *Atmos. Environ.,* 203:18–27.

Millner, P.D. (2009). Bioaerosols associated with animal production operations. *Bioresour. Technol.,* 100(22):5379–5385.

Mohri, M., Rostamizadeh, A., and Talwalkar, A. (2013). Foundations of Machine Learning*, Journal of Chemical Information and Modeling*. MIT Press.

Molina, R.T., López, F.G., Rodríguez, A.M., and Palaciso, I.S. (1996). Pollen production in anemophilous trees. *Grana,* 35(1):38-46f.

Morris, C.E., Sands, D.C., Bardin, M., Jaenicke, R., Vogel, B., Leyronas, C., Ariya, P.A., and Psenner, R. (2011). Microbiology and atmospheric processes: Research challenges concerning the impact of airborne micro-organisms on the atmosphere and climate. *Biogeosciences,* 8(1):17–25.

Mullins, J. and Emberlin, J. (1997). Sampling pollens. *J. Aerosol Sci.,* 28(3):365–370.

Murtagh, F. and Contreras, P. (2017). Algorithms for hierarchical clustering: an overview, II. *WIREs Data Min. Knowl. Discov.,* 7(6):e1219.

Nyomora, A.M.S., Brown, P.H., Pinney, K., and Polito, V.S. (2019). Foliar Application of Boron to Almond Trees Affects Pollen Quality. *J. Am. Soc. Hortic. Sci.,* 125(2):265–270.

O'Connor, D., Healy, D., Hellebust, S., Buters, J., and Sodeau, J. (2014a). Using the WIBS-4 (Waveband Integrated Bioaerosol Sensor) Technique for the On-Line Detection of Pollen Grains. *Aerosol Sci. Technol.,* 48(4):341–349.

O'Connor, D., Lovera, P., Iacopino, D., O'Riordan, A., Healy, D. a., and Sodeau, J.R. (2014b). Using spectral analysis and fluorescence lifetimes to discriminate between grass and tree pollen for aerobiological applications. *Anal. Methods,* 6(6):1633.

O'Connor, D.J., Iacopino, D., Healy, D.A., O'Sullivan, D., and Sodeau, J.R. (2011). The intrinsic fluorescence spectra of selected pollen and fungal spores. *Atmos. Environ.,* 45(35):6451–6458.

Oteros, J. (2019). The role of pollen as a bacterial vector. *J. Environ. Health,* 19:122.

Oteros, J., Pusch, G., Weichenmeier, I., Heimann, U., Möller, R., Röseler, S., Traidl-Hoffmann, C., Schmidt-Weber, C., and Buters, J. (2015). Automatic and online pollen monitoring. *Int. Arch. Allergy Immunol.,* 167(3):158–166.

Pacini, E. and Hesse, M. (2005). Pollenkitt - Its composition, forms and functions. *Flora Morphol. Distrib. Funct. Ecol. Plants,* 200(5):399–415.

Pan, Y. Le, Hill, S.C., Pinnick, R.G., House, J.M., Flagan, R.C., and Chang, R.K. (2011). Dual-excitation-wavelength fluorescence spectra and elastic scattering for differentiation of single airborne pollen and fungal particles. *Atmos. Environ.,* 45(8):1555–1563.

Pan, Y. Le, Huang, H., and Chang, R.K. (2012). Clustered and integrated fluorescence spectra from single atmospheric aerosol particles excited by a 263- and 351-nm laser at New Haven, CT, and Adelphi, MD. *J. Quant. Spectrosc. Radiat. Transf.,* 113(17):2213–2221.

Pan, Y.Y. Le, Hartings, J., Pinnick, R.R.G., Hill, S.S.C., Halverson, J., and Chang, R.K.

(2003). Single-Particle Fluorescence Spectrometer for Ambient Aerosols. *Aerosol Sci. Technol.*, 37(8):628–639.

Pan, Y., Pinnick, R., Hill, S., and Rosen, J. (2007). Single-particle laser-induced-fluorescence spectra of biological and other organic-carbon aerosols in the atmosphere: Measurements at New Haven, Connecticut,. *J.*, 112(24).

Pauling, A., Rotach, M., Gehrig, R., Clot, B., Jäger, S., Cerny, M., Schmidt, R., Bortenschlager, S., Brosch, U., Zwander, H., Schmidt, R., Bobek, M., Litschauer, R., Schantl, H., Koll, H., Langanger, M., Detandt, M., Rybnicek, O., Bergmann, K.C., Jankiewicz, P., Wachter, R., Sommer, J., Díaz de la Guardia Guerrero, C., Alba Sánchez, F., Gutiérrez Bustillo, M., Belmonte, J., Aguinagalde Aizpurua, X., Candau Fernández, P., Moreno Grau, S., Elvira Rendueles, B., Pérez Badía, R., Galán Soldevilla, C., Fernández Casado, M.A., Nava, H.S., Javier Suárez Pérez, F., Hidalgo Fernández, P., Ruiz Valenzuela, L., Fernández González, D., M. Valencia Barrera, R., Vega Maray, A., Jesus Aira, M., Rodríguez-Rajo, J., Mar Trigo Pérez, M., Boi, M., Llorens, L., Mateu Andrés, I., Tortajada, B., Bermejo, D., Thibaudon, M., Hunt, C., Watkeys, L., Clewlow, Y., Ramsay, G., Smith, M., Hrga, I., Páldy, A., Juhasz, M., Travaglini, A., Bucher, E., Mandrioli, P., De Nuntiis, P., Hentges, F., Milkovska, S., Schoenmakers, C.H.H., de Graaf, A.O., de Weger, L., Rapiejko, P., Myszkowska, D., Lipiec, A., Majkowska-Wojciechow, B., Weryszko-Chmielewska, E., Puc, M., Mitrovic Josipovic, M., Dedierv Ljubicic, A., Stamenkovic, D., Sikoparija, B., Dahl, Å., Kofol-Seliger, A., and Micieta, K. (2012). A method to derive vegetation distribution maps for pollen dispersion models using birch as an example. *Int. J. Biometeorol.*, 56(5):949–958.

Peden, D. and Reed, C.E. (2010). Environmental and occupational allergies. *J. Allergy Clin. Immunol.*,.

Percy, D.F. and Everitt, B.S. (2006). Cluster Analysis (3rd Edition). *J. Oper. Res. Soc.*,.

Perring, A., Schwarz, J., Baumgardner, D., Hernandez, M., Spracklen, D., Heald, C., Gao, R., Kok, G., McMeeking, G., McQuaid, J., and Fahey, D. (2015). Airborne observations of regional variation in fluorescent aerosol across the United States. *J. Geophys. Res. Atmos.*, 120(3):1153–1170.

Pinnick, R.G., Fernandez, E., Rosen, J.M., Hill, S.C., Wang, Y., and Pan, Y.L. (2013). Fluorescence spectra and elastic scattering characteristics of atmospheric aerosol in Las Cruces, New Mexico, USA: Variability of concentrations and possible constituents and sources of particles in various spectral clusters. *Atmos. Environ.*, 65:195–204.

Pinnick, R.G., Hill, S.C., Pan, Y. Le, and Chang, R.K. (2004). Fluorescence spectra of atmospheric aerosol at Adelphi, Maryland, USA: Measurement and classification of single particles containing organic carbon. *Atmos. Environ.,* 38(11):1657–1672.

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.

Pöhlker, C., Huffman, J.A., and Pöschl, U. (2012). Autofluorescence of atmospheric bioaerosols - Fluorescent biomolecules and potential interferences. *Atmos. Meas. Tech.,* 5(1):37–71.

Pöschl, U. (2005). Atmospheric aerosols: Composition, transformation, climate and health effects. *Angew. Chemie - Int. Ed.,*.

Prabhat, P., Ram, S., Sally Ward, E., and Ober, R.J. (2004). Simultaneous imaging of different focal planes in fluorescence microscopy for the study of cellular dynamics in three dimensions. *IEEE Trans. Nanobioscience,* 3(4):237–242.

Ranzato, M., Taylor, P.E., House, J.M., Flagan, R.C., LeCun, Y., and Perona, P. (2007). Automatic recognition of biological particles in microscopic images. *Pattern Recognit. Lett.,* 28(1):31–39.

Rapiejko, P., Szczygielski, K., Jurkiewicz, D., and Stankiewicz, W. (2007). Threshold pollen count necessary to evoke allergic symptoms. *Otolaryngol. Pol.,* 61(4):591–594.

Rebotier, T.P. and Prather, K.A. (2007). Aerosol time-of-flight mass spectrometry data analysis: A benchmark of clustering algorithms. *Anal. Chim. Acta,*.

Rendón, E., Abundez, I., Arizmendi, A., and Quiroz, E.M. (2011). Internal versus External cluster validation indexes. *Int. J.,* 5(1):27–34.

Robinson, N., Allan, J., Huffman, J., Kaye, P., Foot, V., and Gallagher, M. (2013). Cluster analysis of WIBS single-particle bioaerosol data. *Atmos. Meas. Tech.,* 6(2):337–347.

Rodriguez-Damian, M., Cernadas, E., Formella, A., Fernandez-Delgado, M., and Pilar De Sa-Otero (2006). Automatic detection and classification of grains of pollen based on shape and texture. *IEEE Trans. Syst. Man Cybern. Part C (Applications Rev.,* 36(4):531–542.

Rokach, L. (2010). Ensemble-based classifiers. *Artif. Intell. Rev.,* 33(1–2):1–39.

RStudio Team, - (2016). RStudio: Integrated Development for R. *[Online] RStudio, Inc., Boston, MA URL http//www. rstudio. com,* RStudio, Inc., Boston, MA.

Runions, C.J. and Owens, J.N. (2002). Sexual Reproduction of Interior Spruce (Pinaceae). I. Pollen Germination to Archegonial Maturation. *Int. J. Plant Sci.,* 160(4):631–640.

Ruske, S., Topping, D., Foot, V., Kaye, P., Stanley, W., Crawford, I., Morse, A., and Gallagher, M. (2017). Evaluation of machine learning algorithms for classification of primary biological aerosol using a new UV-LIF spectrometer. *Atmos. Meas. Tech.,* 10(2):695–708.

Ruske, S., Topping, D.O., Foot, V.E., Morse, A.P., and Gallagher, M.W. (2018). Machine learning for improved data analysis of biological aerosol using the WIBS. *Atmos. Meas. Tech.,* 11(11):6203–6230.

Saari, S., Reponen, T., and Keskinen, J. (2014). Performance of Two Fluorescence-Based Real-Time Bioaerosol Detectors: BioScout vs. UVAPS. *Aerosol Sci. Technol.,* 48(4):371–378.

Sachkov, M., Shustov, B., Savanov, I., and Gómez de Castro, A.I. (2014). WSO-UV project for high-resolution spectroscopy and imaging. *Astron. Nachrichten,* 335(1):46–50.

Sahoo, H. (2012). Fluorescent labeling techniques in biomolecules: A flashback. *RSC Adv.,* 2(18):7017–7029.

Sassen, K. (2008). Boreal tree pollen sensed by polarization lidar: Depolarizing biogenic chaff. *Geophys. Res. Lett.,* 35(18).

Sato, S. and Peet, M.M. (2005). Effects of moderately elevated temperature stress on the timing of pollen release and its germination in tomato (Lycopersicon esculentum Mill.). *J. Hortic. Sci. Biotechnol.,* 80(1):23–28.

Satterwhite, M.B. (1990). Spectral Luminescence Of Plant Pollen., *in 10th Annual International Symposium on Geoscience and Remote Sensing,* pp. 1945–1948.

Šaulienė, I., Šukienė, L., Daunys, G., Valiulis, G., Vaitkevičius, L., Matavulj, P., Brdar, S., Panic, M., Sikoparija, B., Clot, B., Crouzy, B., and Sofiev, M. (2019). Automatic pollen recognition with the Rapid-E particle counter: the first-level procedure, experience and next steps. *Atmos. Meas. Tech. Discuss.,* 2:1–33.

Sauvageat, E., Yanick, Z., Tummon, F., Clot, C., and Courzy, B. (2019). No Title. *Prep,*.

Savage, N. and Huffman, J.A. (2018). Evaluation of a hierarchical agglomerative clustering method applied to WIBS laboratory data for improved discrimination of biological particles by comparing data preparation techniques. *Atmos. Meas. Tech.,* 11(8):4929–4942.

Savage, N., Krentz, C., Könemann, T., Han, T., Mainelis, G., Pöhlker, C., and Huffman, J.A. (2017). Systematic Characterization and Fluorescence Threshold Strategies for the Wideband Integrated Bioaerosol Sensor (WIBS) Using Size-Resolved Biological and Interfering Particles. *Atmos. Meas. Tech.,* 10(11):4279–4302.

Schoper, J.B., Lambert, R.J., and Vasilas, B.L. (2010). Pollen Viability, Pollen Shedding, and Combining Ability for Tassel Heat Tolerance in Maize 1. *Crop Sci.,* 27(1):27–31.

Schwendemann, A., Wang, G., Mertz, M., McWilliams, R., Thatcher, S., and Osborn, J. (2007). Aerodynamics of saccate pollen and its implications for wind pollination. *Am. J. Bot.,* 94(8):1371–1381.

Sibson, R. (1981). A Brief Description of Natural Neighbour Interpolation.*, in Interpreting Multivariate Data,* p. 374.

Sivaprakasam, V., Lin, H., Huston, A.L., and Eversole, J.D. (2011). Spectral characterization of biological aerosol particles using two-wavelength excited laser-induced fluorescence and elastic scattering measurements. *Opt. Express,* 19(7):6191–6208.

Smith, T.T. (1922). Spherical aberration in thin lenses. *Phys. Rev.,* 19(3):276–277.

Sodeau, J.R. and O’Connor, D.J. (2016). Bioaerosol Monitoring of the Atmosphere for Occupational and Environmental Purposes. *Compr. Anal. Chem.,* 73:391–420.

Sofiev and Bergmann (2013). Allergenic Pollen*, Allergenic Pollen*.

SONY (n.d.). ICX205 AL Fact Sheet. Available at https://www.1stvision.com/cameras/sensor_specs/ICX205.pdf

Spieksma, F. (1990). Pollinosis in Europe: New observations and developments. *Rev. Palaeobot. Palynol.,* 64(1–4):35–40.

Stach, A., Smith, M., Prieto Baena, J.C., and Emberlin, J. (2008). Long-term and short-term forecast models for Poaceae (grass) pollen in Poznań, Poland, constructed using regression analysis. *Environ. Exp. Bot.,* 62(3):323–332.

Stacy, E.A., Hamrick, J.L., Nason, J.D., Hubbell, S.P., Foster, R.B., and Condit, R. (2002). Pollen Dispersal in Low-Density Populations of Three Neotropical Tree Species. *Am. Nat.,* 148(2):275–298.

Staff, I.A., Taylor, P.E., Smith, P., Singh, M.B., and Knox, R.B. (1990). Cellular localization of water soluble, allergenic proteins in rye-grass (Lolium perenne) pollen using monoclonal and specific IgE antibodies with immunogold probes.

Histochem. J., 22(5):276–290.

Stanghellini, L., Shaw, R.A., Mutchler, M., Palen, S., Balick, B., and Blades, J.C. (2002). Optical Slitless Spectroscopy of Large Magellanic Cloud Planetary Nebulae: A Study of the Emission Lines and Morphology. *Astrophys. J.*, 575(1):178.

Strobl, C., Boulesteix, A.-L., Zeileis, A., and Hothorn, T. (2007). Bias in random forest variable importance measures: illustrations, sources and a solution. *BMC Bioinformatics*, 8(1):25.

Swanson, B. and Huffman, J.A. (2019). Pollen Clustering Strategies Using a Newly Developed Single-Particle Fluorescence Spectrometer. *Aerosol Sci. Technol.*, In Review.

Swanson, B. and Huffman, J.A. (2018). Development and characterization of an inexpensive single-particle fluorescence spectrometer for bioaerosol monitoring. *Opt. Express*, 26(3):3646–3660.

Taillard, É.D. (2003). Heuristic methods for large centroid clustering problems. *J. Heuristics*, 9(1):51–73.

Tanaka, N., Uehara, K., and Murata, J. (2004). Correlation between pollen morphology and pollination mechanisms in the Hydrocharitaceae. *J. Plant Res.*, 117(4):265–276.

Taylor, P.E., Flagan, R.C., Miguel, A.G., Valenta, R., and Glovsky, M.M. (2004). Birch pollen rupture and the release of aerosols of respirable allergens. *Clin. Exp. Allergy*, 34(10):1591–1596.

Taylor, P.E., Jacobson, K.W., House, J.M., and Glovsky, M.M. (2007). Links between pollen, atopy and the asthma epidemic. *Int. Arch. Allergy Immunol.*, 144(2):162–170.

Tello-Mijares, S. and Flores, F. (2016). A Novel Method for the Separation of Overlapping Pollen Species for Automated Detection and Classification. *Comput. Math. Methods Med.*,.

Tseng, Y.-T. and Kawashima, S. (2019). Applying a pollen forecast algorithm to the Swiss Alps clarifies the influence of topography on spatial representativeness of airborne pollen data. *Atmos. Environ.*, 212:153–162.

USDA (2016). PLANTS Database | USDA PLANTS. *Plant Guid.*,.

Vrtala, S., Grote, M., Duchêne, M., Vanree, R., Kraft, D., Scheiner, O., and Valenta, R. (1993). Properties of tree and grass pollen allergens: Reinvestigation of the linkage between solubility and allergenicity. *Int. Arch. Allergy Immunol.*, 102(2):160–169.

Wayne, P., Foster, S., Connolly, J., Bazzaz, F., and Epstein, P. (2002). Production of allergenic pollen by ragweed (Ambrosia artemisiifolia L.) is increased in CO2-enriched atmospheres. *Ann. Allergy, Asthma Immunol.,* 88(3):279–282.

Weber, M. and Ulrich, S. (2017). PalDat 3.0–second revision of the database, including a free online publication tool. *Grana,*.

Weber, R.W. (2010). Pollen Identification. *Ann. Allergy, Asthma Immunol.,* 80(2):141–148.

Wéry, N. (2014). Bioaerosols from composting facilities-a review. *Front. Cell. Infect. Microbiol.,* (4):42.

Williams, C.G. and Després, V. (2017). Northern Hemisphere forests at temperate and boreal latitudes are substantial pollen contributors to atmospheric bioaerosols. *For. Ecol. Manage.,* 401:187–191.

Wlodarski, M., Kaliszewski, M., Kwasny, M., Kopczynski, K., Zawadzki, Z., Mierczyk, Z., Mlynczak, J., Trafny, E., and Szpakowska, M. (2006). Fluorescence excitation-emission matrices of selected biological materials., *in Optically Based Biological and Chemical Detection for Defence III,* p. 639806.

Wopfner, N., Gadermaier, G., Egger, M., Asero, R., Ebner, C., Jahn-Schmid, B., and Ferreira, F. (2005). The spectrum of allergens in ragweed and mugwort pollen. *Int. Arch. Allergy Immunol.,* 138(4):337–346.

Wu, Y., Calis, A., Luo, Y., Chen, C., Lutton, M., Rivenson, Y., Lin, X., Koydemir, H.C., Zhang, Y., Wang, H., Göröcs, Z., and Ozcan, A. (2018). Label-Free Bioaerosol Sensing Using Mobile Microscopy and Deep Learning. *ACS Photonics,* 5(11):4617–4627.

Xiaowei Xu, Ester, M., Kriegel, H.-P., and Sander, J. (1998). A distribution-based clustering algorithm for mining in large spatial databases., *in Proceedings 14th International Conference on Data Engineering,* pp. 324–331.

Xiong, B., Zhou, R., Hao, J., Jia, Y., He, Y., and Yeung, E.S. (2013). Highly sensitive sulphide mapping in live cells by kinetic spectral analysis of single Au-Ag core-shell nanoparticles. *Nat. Commun.,* 4:1708.

Zhang, P., Zhao, Y., Liao, X., Yang, W., Zhu, Y., and Huang, H. (2013). Development and calibration of a single UV LED based bioaerosol monitor. *Opt. Express,* 21(22):26303–10.

Zhu, W., Liu, Q., and Wu, Y. (2015). Aerosol absorption measurement at SWIR with water vapor interference using a differential photoacoustic spectrometer. *Opt.*

Express, 23(18):23108–23116.

Ziska, L.H. and Caulfield, F.A. (2002). Rising CO2 and pollen production of common ragweed (Ambrosia artemisiifolia L.), a known allergy-inducing species: implications for public health. *Funct. Plant Biol.*, 27(10):893–898.

Appendix A: Chapter 4 Supplement

Supplementary information for Swanson, Benjamin E. Swanson, and J. Alex Huffman. " Pollen Clustering Strategies Using a Newly Developed Single-Particle Fluorescence Spectrometer." Aerosol Science and Technology, 2019 (In Review)

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) and Benjamin E. Swanson (Benjamin.swanson@du.edu)

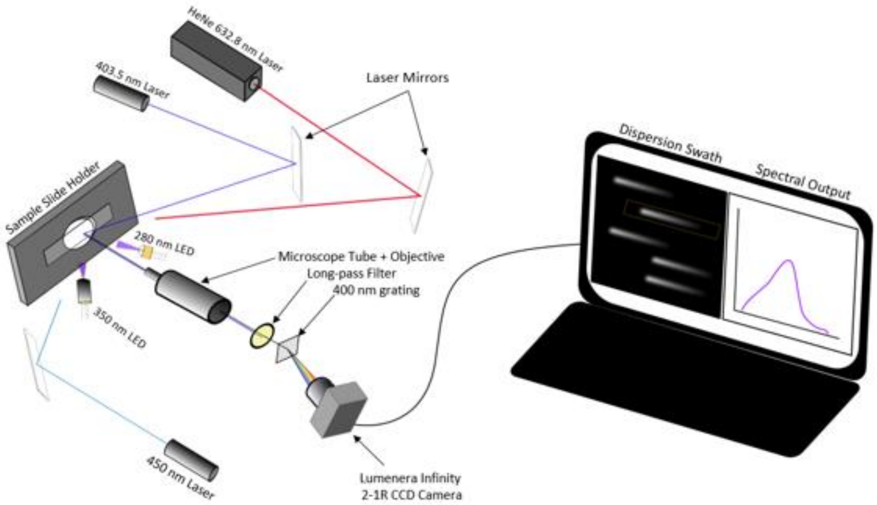


Figure A1: Schematic of instrumental design and operation.

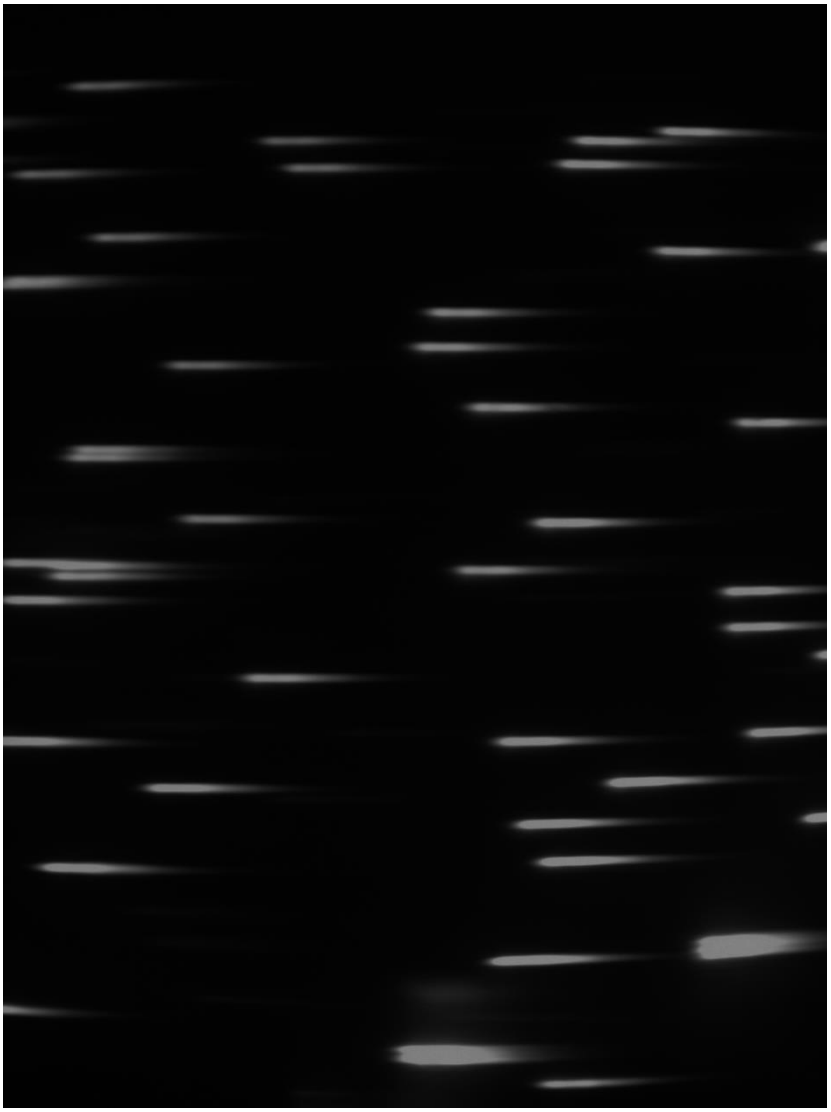


Figure A2: Camera viewing area with approximately 50 visible particle signals, represented as dispersed swaths of ~400 to 700 nm fluorescent light (left to right for each swath)

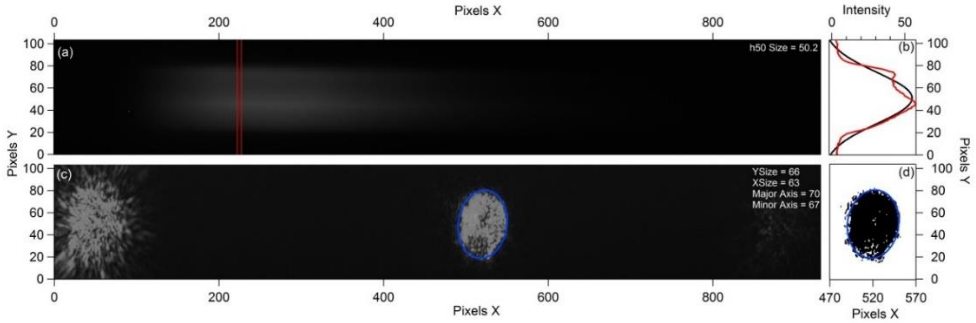


Figure A3: Particle sizing Figure, associated with Figure 1 in the main text (where images in (a) and (c) were chopped for visual clarity).

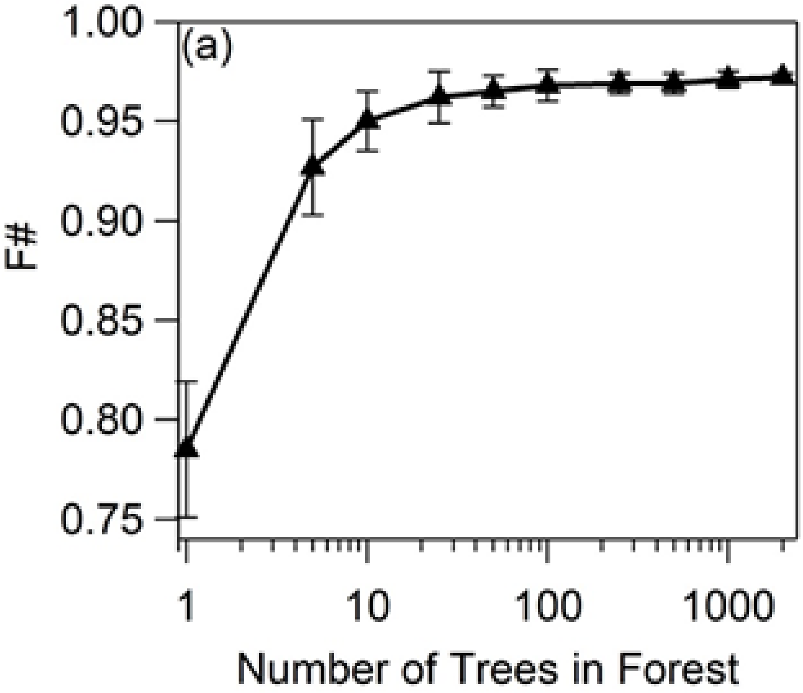


Figure A4: Analysis of the Random Forest model accuracy as the number of trees are increased in the analysis. The triangle represents the average of 5 trials, and the bars represent the deviation.

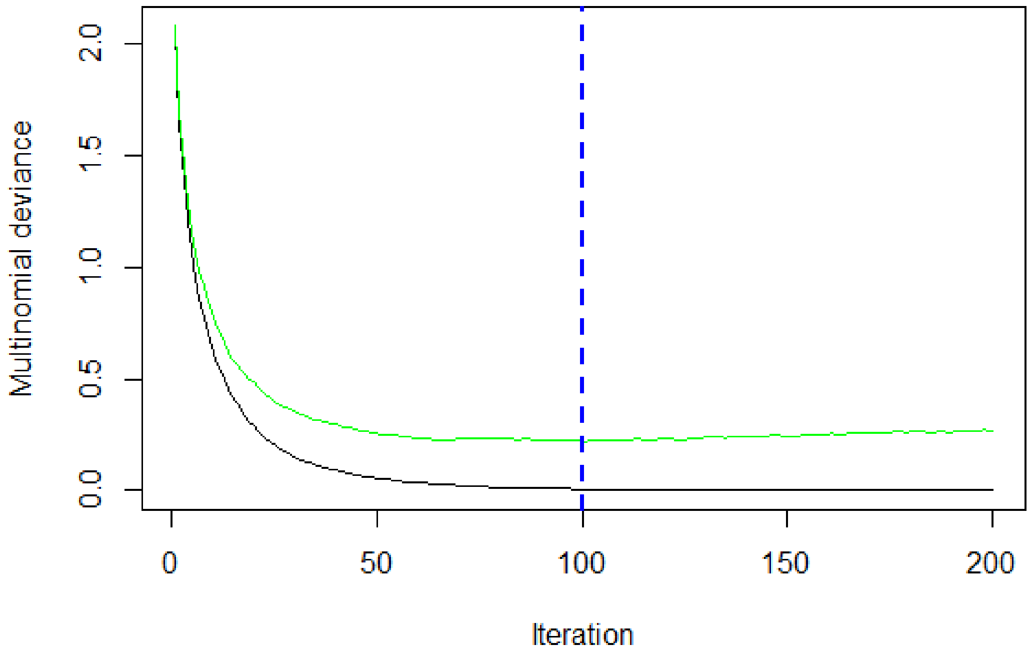


Figure A5: Exponential loss plot for GB model. Black trace indicates the reduction of error as subsequent trees are developed. Green trace represents reduction in error after *k* folds of the cross-validation test sets. Blue dotted line indicates minima on green trace.

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

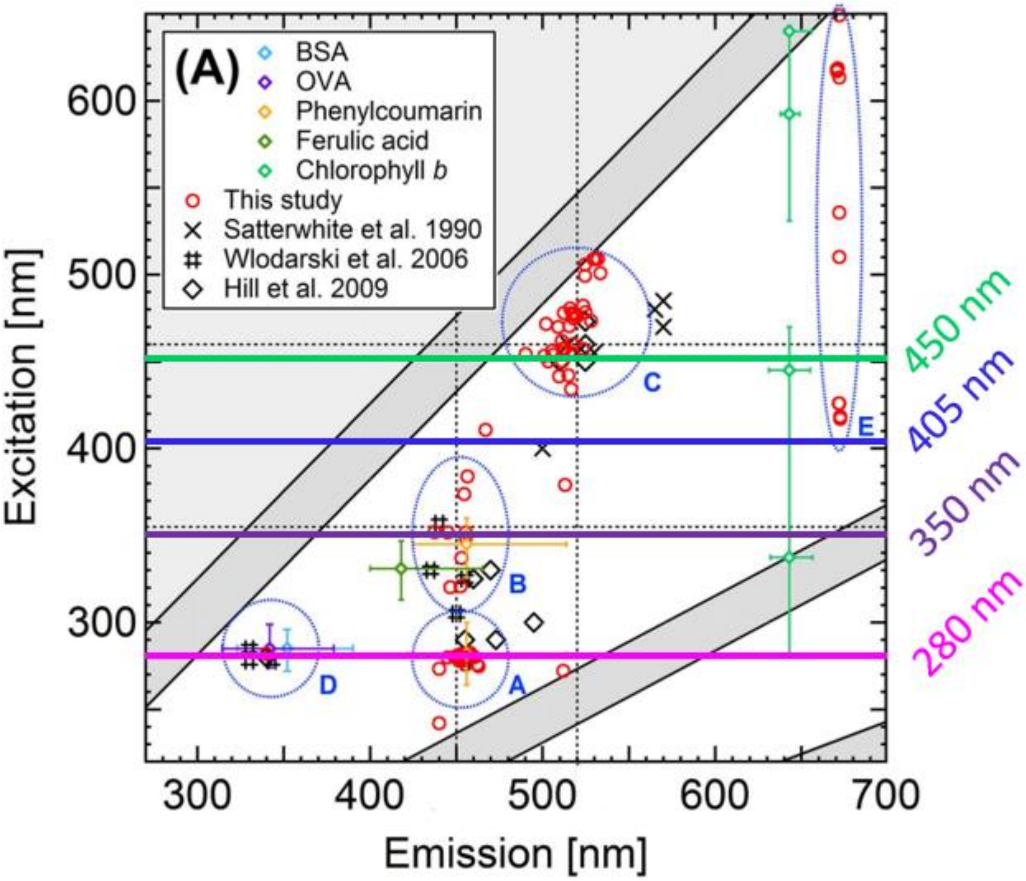


Figure A6: Previous work on pollen excitation and emission variables, with the excitation waves from this instrument shown as horizontal colored lines. Adapted from Pöhlker et al., 2013.

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.

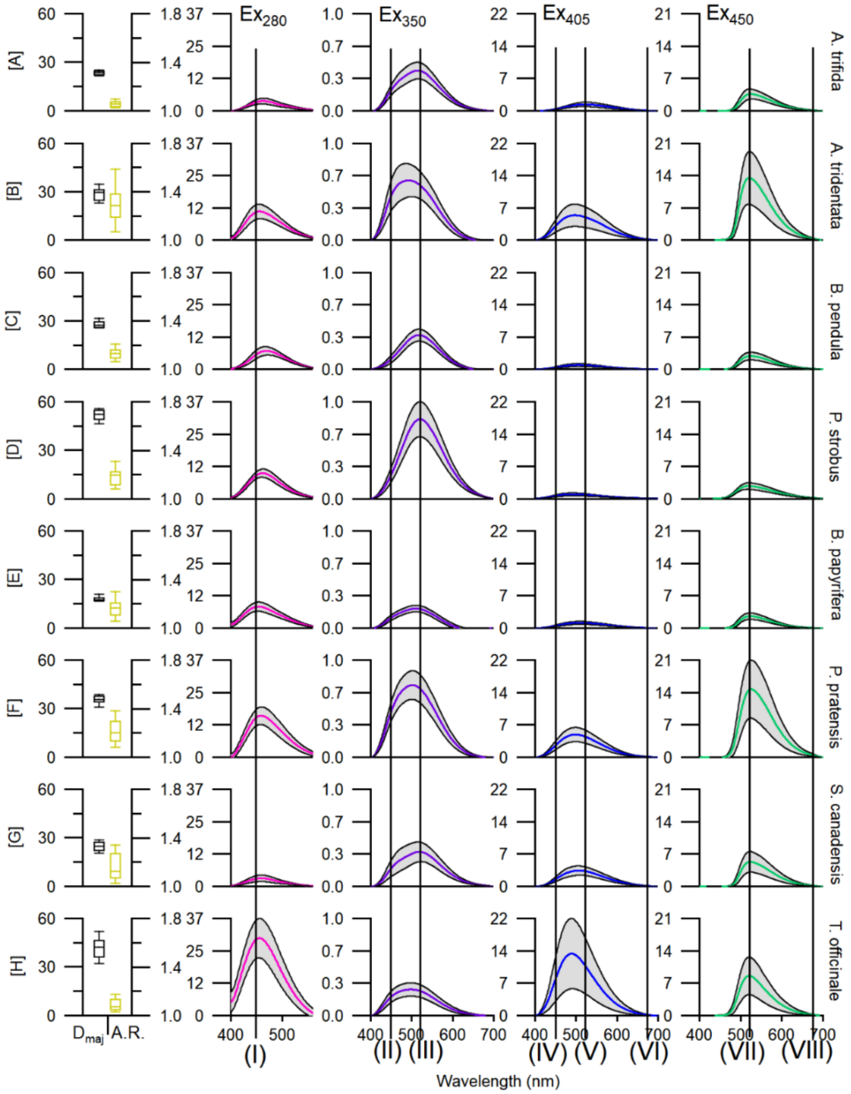


Figure A7: Particle size and spectral characteristics of the eight pollen species examined. Analogous to Figure 3 in main text, but with black vertical lines added to represent fluorophores emission modes as defined in Section 3.1.

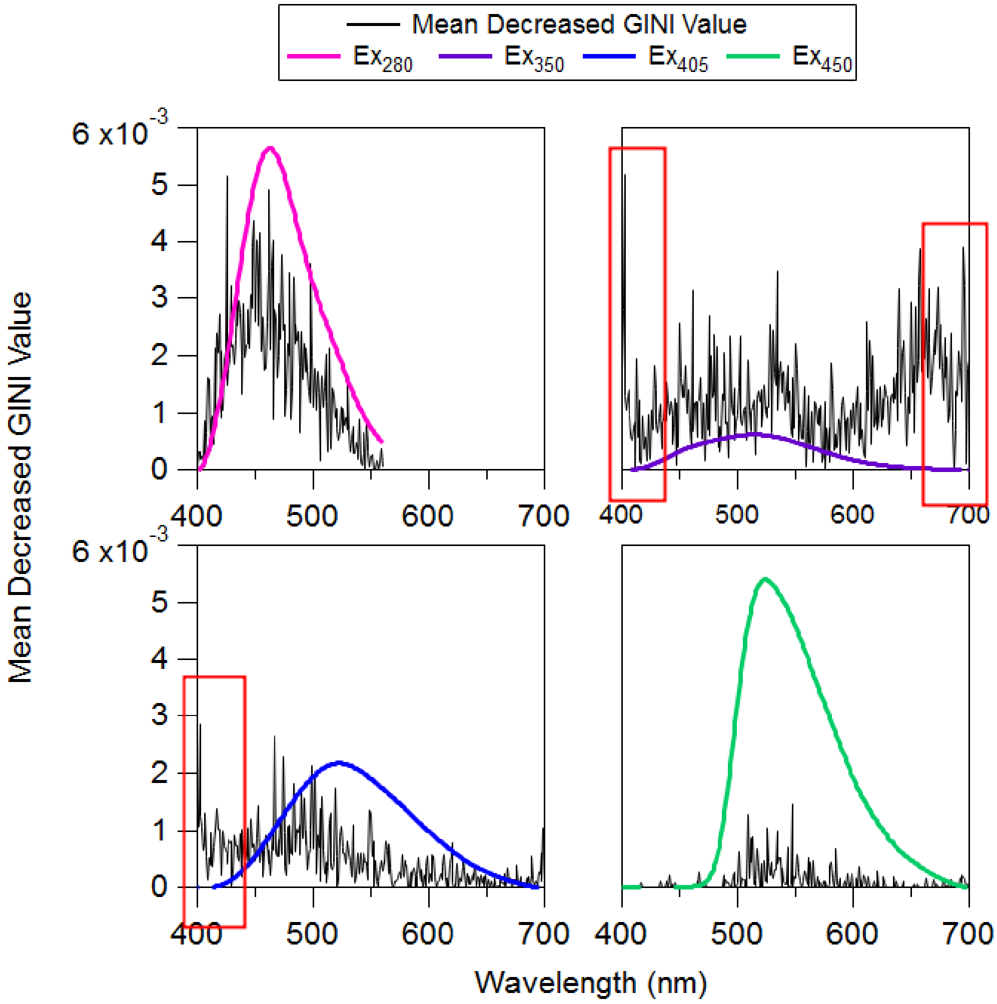


Figure A8: Model importance values (black), before spectra are reduced, plotted against the reference spectra (color) shown in Figure 5. Areas with appreciably high importance corresponding to noise are boxed in red. Emission data was reduced roughly outside the red boxed areas to where the importance curve trends high and the spectral curves lower.

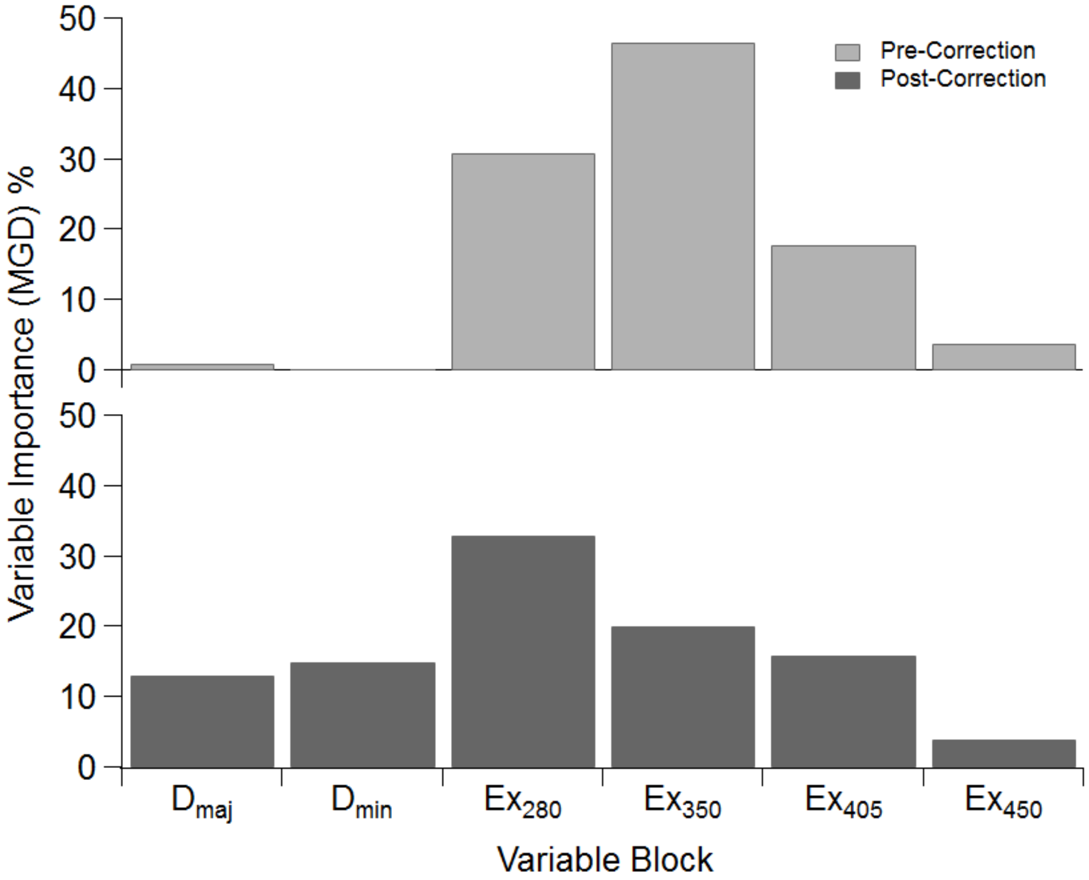


Figure A9: Importance changes after spectral reduction and size weighting in the gradient boosting system.

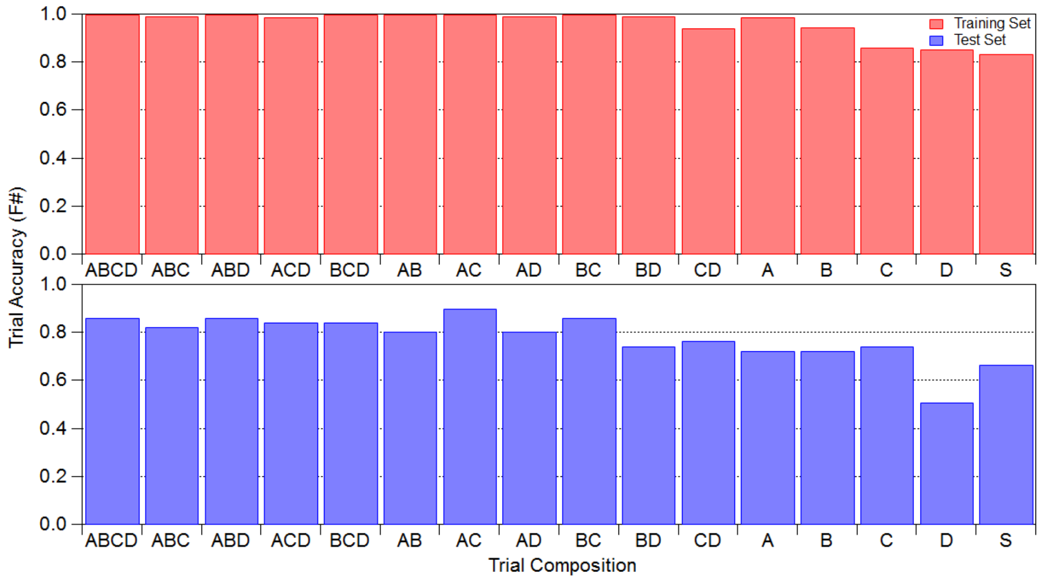


Figure A10: Accuracy of the RF algorithm following fifteen combinations of input variables. Excitation sources represented here as (A) 280 nm, (B) 350 nm, (C), 405 nm, and (D) 450 nm. All trials consist of a subset of 25% of the particle spectral data predicted to training models from 75% of the data. The final column represents no emission data, but only sizing (S) variables.

Appendix B: Chapter 5 Supplement

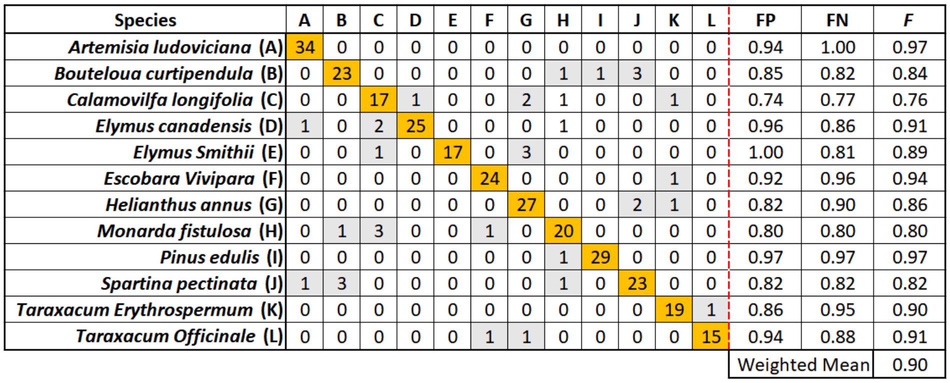
Supplementary information for Benjamin E. Swanson, Samir Rezgui, and J. Alex Huffman. “Pollen Classification Using a Newly Developed Fluorescence Spectrometer.” For submission to Aerobiologia (in Prep).

Pollen Classification with a Recently Developed Fluorescence Spectrometer

Benjamin E. Swanson and J. Alex Huffman

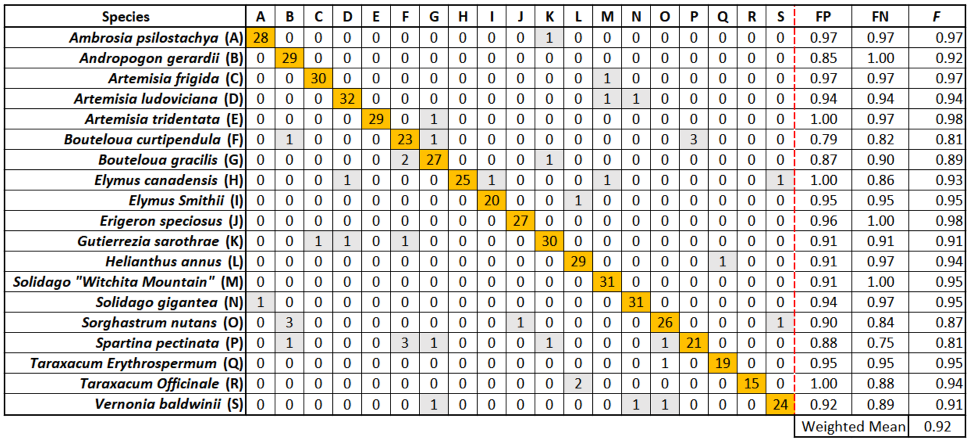
University of Denver, Department of Chemistry and Biochemistry, Denver CO 80210, USA

Correspondence to: J. Alex Huffman (alex.huffman@du.edu) and Benjamin E. Swanson (Benjamin.swanson@du.edu)



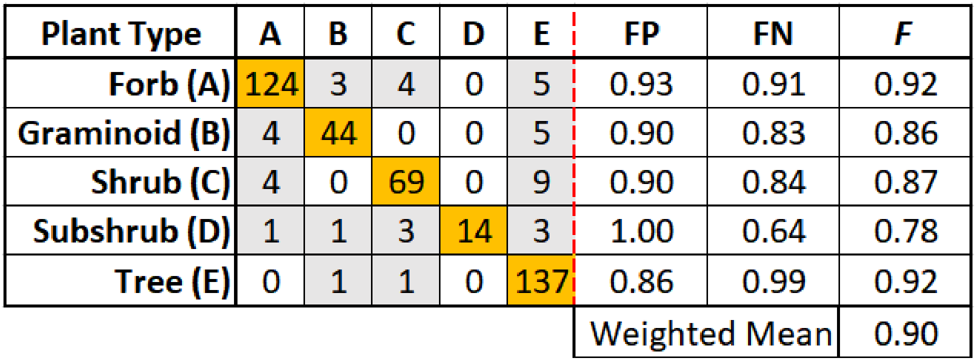
Species	A	B	C	D	E	F	G	H	I	J	K	L	FP	FN	*F*
***Artemisia ludoviciana* (A)**	34	0	0	0	0	0	0	0	0	0	0	0	0.94	1.00	0.97
***Bouteloua curtipendula* (B)**	0	23	0	0	0	0	0	1	1	3	0	0	0.85	0.82	0.84
***Calamovilfa longifolia* (C)**	0	0	17	1	0	0	2	1	0	0	1	0	0.74	0.77	0.76
***Elymus canadensis* (D)**	1	0	2	25	0	0	0	1	0	0	0	0	0.96	0.86	0.91
***Elymus Smithii* (E)**	0	0	1	0	17	0	3	0	0	0	0	0	1.00	0.81	0.89
***Escobara Vivipara* (F)**	0	0	0	0	0	24	0	0	0	0	1	0	0.92	0.96	0.94
***Helianthus annus* (G)**	0	0	0	0	0	0	27	0	0	2	1	0	0.82	0.90	0.86
***Monarda fistulosa* (H)**	0	1	3	0	0	1	0	20	0	0	0	0	0.80	0.80	0.80
***Pinus edulis* (I)**	0	0	0	0	0	0	0	1	29	0	0	0	0.97	0.97	0.97
***Spartina pectinata* (J)**	1	3	0	0	0	0	0	1	0	23	0	0	0.82	0.82	0.82
***Taraxacum Erythrospermum* (K)**	0	0	0	0	0	0	0	0	0	0	19	1	0.86	0.95	0.90
***Taraxacum Officinale* (L)**	0	0	0	0	0	1	1	0	0	0	0	15	0.94	0.88	0.91
													Weighted Mean		0.90

Table B1. Confusion matrix results from RF classification of the sampling window P3 by Species.



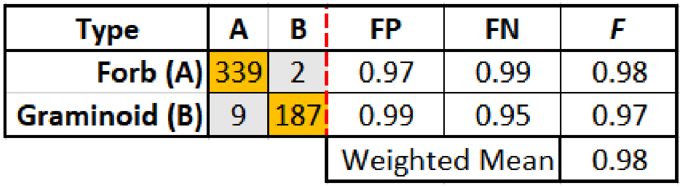
Species	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	FP	FN	*F*
***Ambrosia psilostachya* (A)**	28	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0.97	0.97	0.97
***Andropogon gerardii* (B)**	0	29	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.85	1.00	0.92
***Artemisia frigida* (C)**	0	0	30	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0.97	0.97	0.97
***Artemisia ludoviciana* (D)**	0	0	0	32	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0.94	0.94	0.94
***Artemisia tridentata* (E)**	0	0	0	0	29	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1.00	0.97	0.98
***Bouteloua curtipendula* (F)**	0	1	0	0	0	23	1	0	0	0	0	0	0	0	0	3	0	0	0	0.79	0.82	0.81
***Bouteloua gracilis* (G)**	0	0	0	0	0	2	27	0	0	0	1	0	0	0	0	0	0	0	0	0.87	0.90	0.89
***Elymus canadensis* (H)**	0	0	0	1	0	0	0	25	1	0	0	0	1	0	0	0	0	0	1	1.00	0.86	0.93
***Elymus Smithii* (I)**	0	0	0	0	0	0	0	0	20	0	0	1	0	0	0	0	0	0	0	0.95	0.95	0.95
***Erigeron speciosus* (J)**	0	0	0	0	0	0	0	0	0	27	0	0	0	0	0	0	0	0	0	0.96	1.00	0.98
***Gutierrezia sarothrae* (K)**	0	0	1	1	0	1	0	0	0	0	30	0	0	0	0	0	0	0	0	0.91	0.91	0.91
***Helianthus annus* (L)**	0	0	0	0	0	0	0	0	0	0	0	29	0	0	0	0	1	0	0	0.91	0.97	0.94
***Solidago "Witchita Mountain"* (M)**	0	0	0	0	0	0	0	0	0	0	0	0	31	0	0	0	0	0	0	0.91	1.00	0.95
***Solidago gigantea* (N)**	1	0	0	0	0	0	0	0	0	0	0	0	0	31	0	0	0	0	0	0.94	0.97	0.95
***Sorghastrum nutans* (O)**	0	3	0	0	0	0	0	0	0	1	0	0	0	0	26	0	0	0	1	0.90	0.84	0.87
***Spartina pectinata* (P)**	0	1	0	0	0	3	1	0	0	0	1	0	0	0	1	21	0	0	0	0.88	0.75	0.81
***Taraxacum Erythrospermum* (Q)**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	19	0	0	0.95	0.95	0.95
***Taraxacum Officinale* (R)**	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	15	0	1.00	0.88	0.94
***Vernonia baldwinii* (S)**	0	0	0	0	0	0	1	0	0	0	0	0	0	1	1	0	0	0	24	0.92	0.89	0.91
																				Weighted Mean		0.92

Table B2._Confusion matrix results from RF classification of the sampling window P4 by Species.



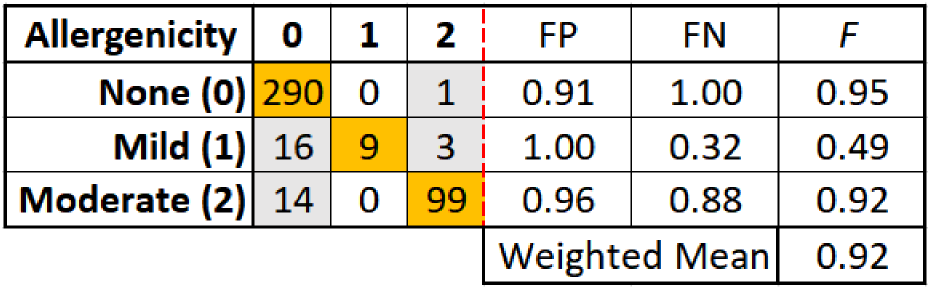
Plant Type	A	B	C	D	E	FP	FN	*F*
Forb (A)	124	3	4	0	5	0.93	0.91	0.92
Graminoid (B)	4	44	0	0	5	0.90	0.83	0.86
Shrub (C)	4	0	69	0	9	0.90	0.84	0.87
Subshrub (D)	1	1	3	14	3	1.00	0.64	0.78
Tree (E)	0	1	1	0	137	0.86	0.99	0.92
						Weighted Mean		0.90

Table B3. Confusion matrix results from RF classification of the sampling window P1 by Plant Type.



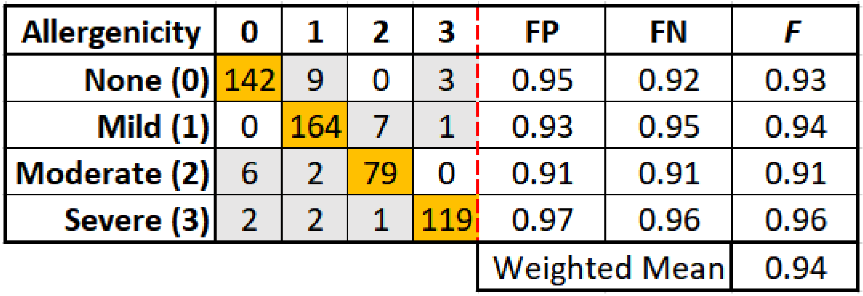
Type	A	B	FP	FN	*F*
Forb (A)	339	2	0.97	0.99	0.98
Graminoid (B)	9	187	0.99	0.95	0.97
			Weighted Mean		0.98

Table B4. Confusion matrix results from RF classification of the sampling window P4 by Plant Type.



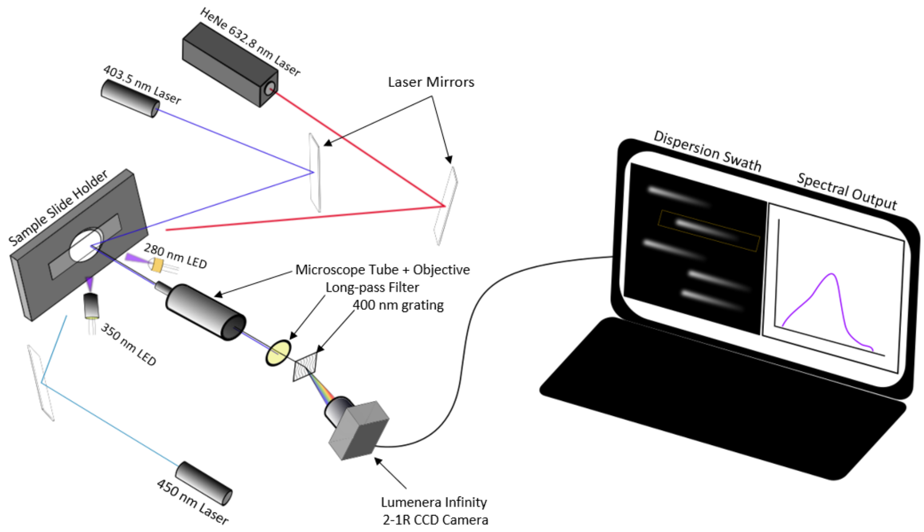
Allergenicity	**0**	**1**	**2**	FP	FN	*F*
None (0)	290	0	1	0.91	1.00	0.95
Mild (1)	16	9	3	1.00	0.32	0.49
Moderate (2)	14	0	99	0.96	0.88	0.92
				Weighted Mean		0.92

Table B5. Confusion matrix results from RF classification of the sampling window P1 by Allergenicity level

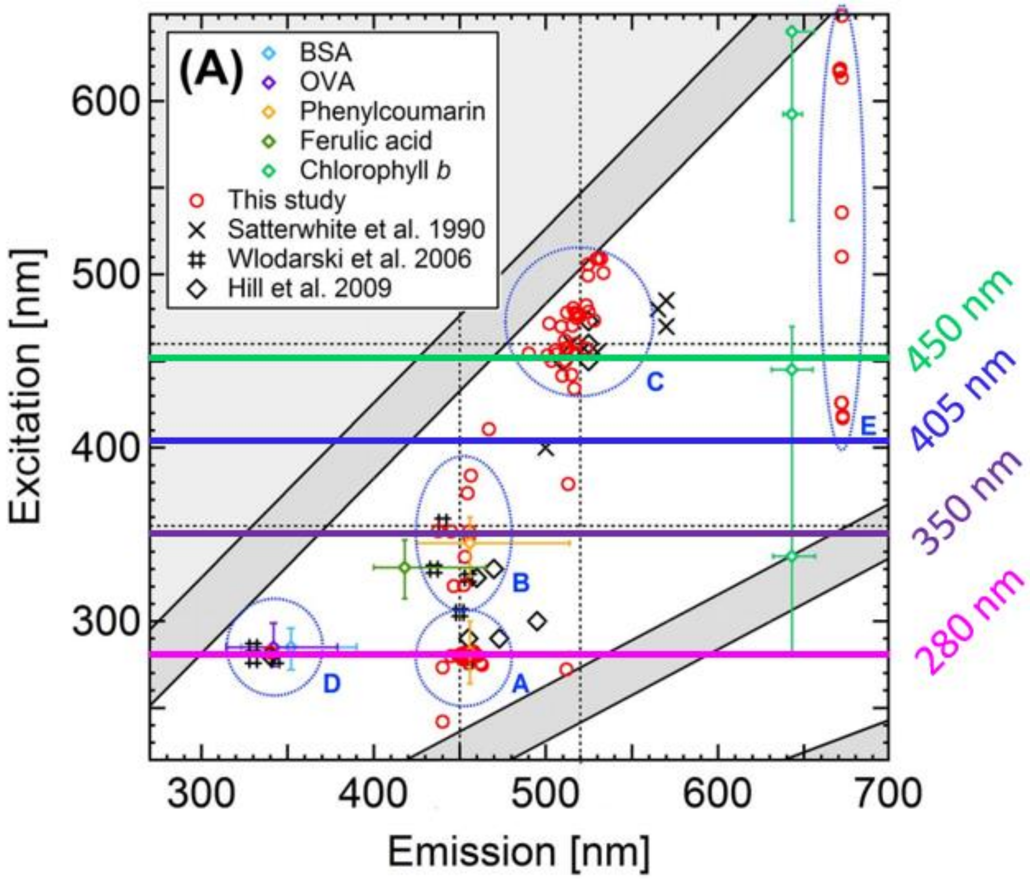


Allergenicity	0	1	2	3	FP	FN	*F*
None (0)	142	9	0	3	0.95	0.92	0.93
Mild (1)	0	164	7	1	0.93	0.95	0.94
Moderate (2)	6	2	79	0	0.91	0.91	0.91
Severe (3)	2	2	1	119	0.97	0.96	0.96
					Weighted Mean		0.94

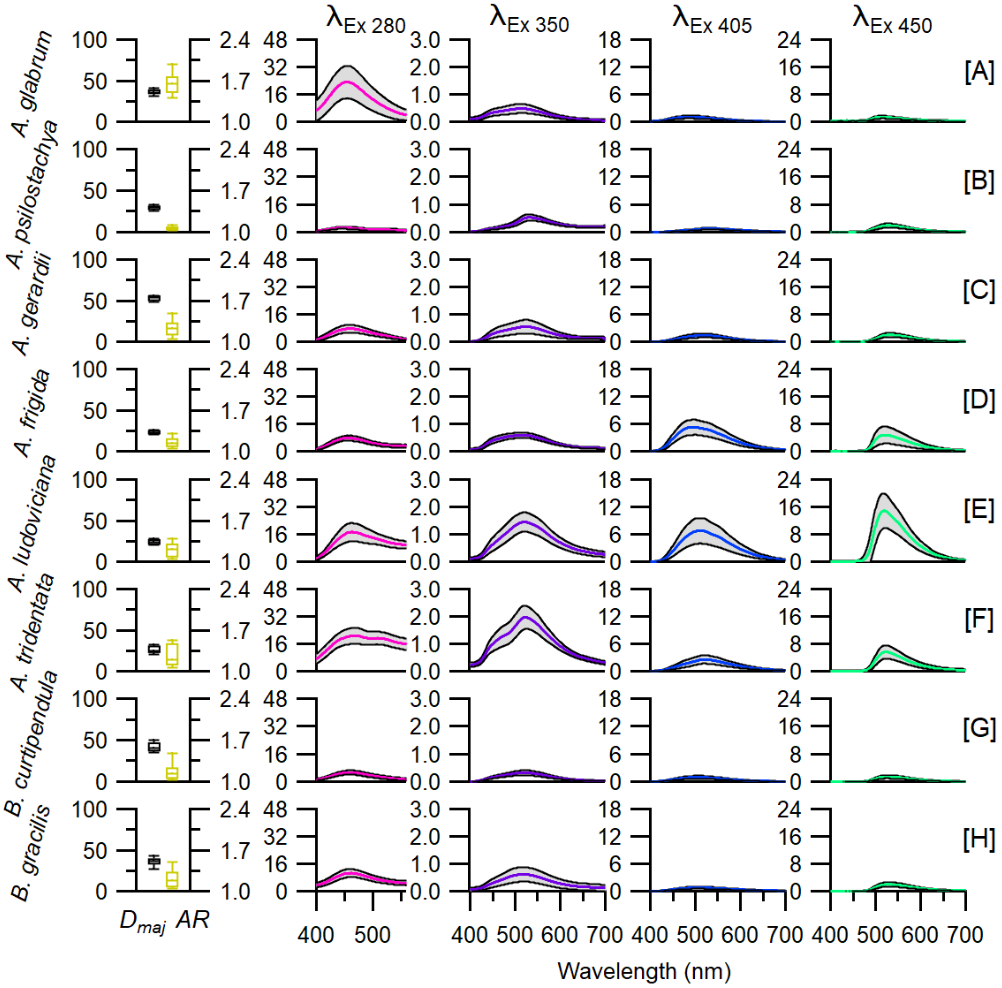
Table B6. Confusion matrix results from RF classification of the sampling window P4 by Allergenicity level



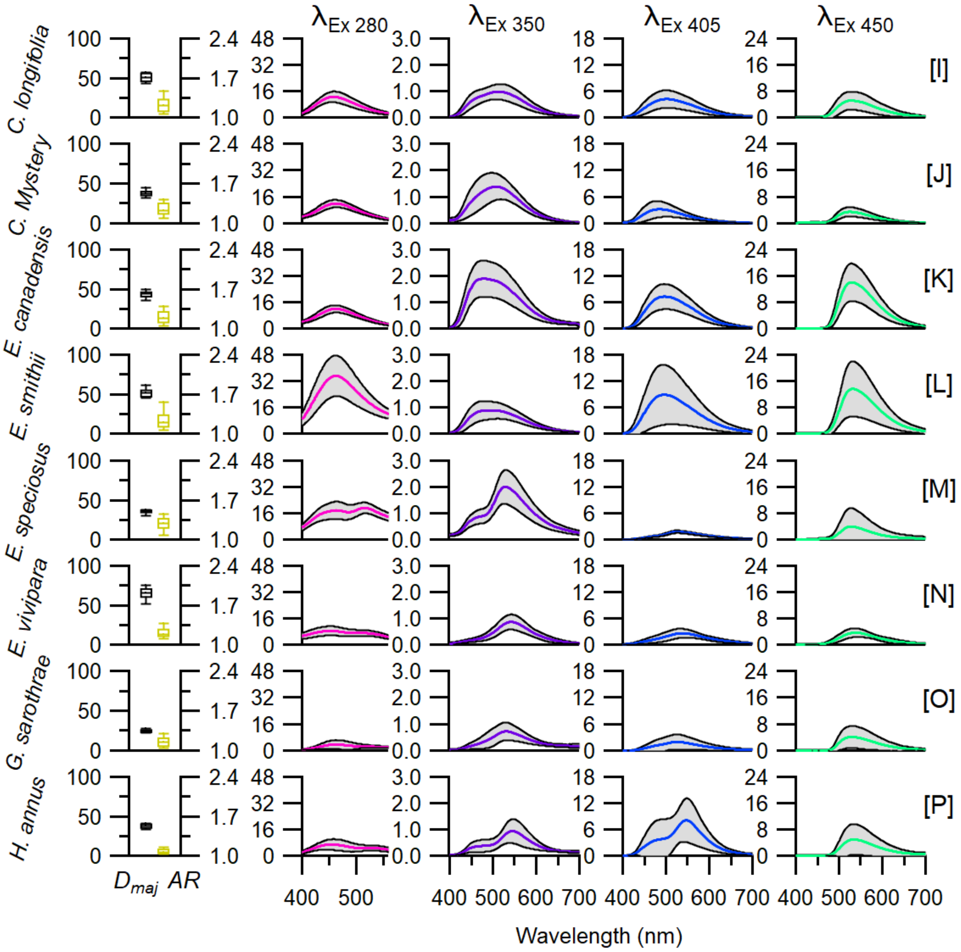
Appendix Figure B1: Technical diagram of the instrument, reproduced from Swanson and Huffman, 2019.



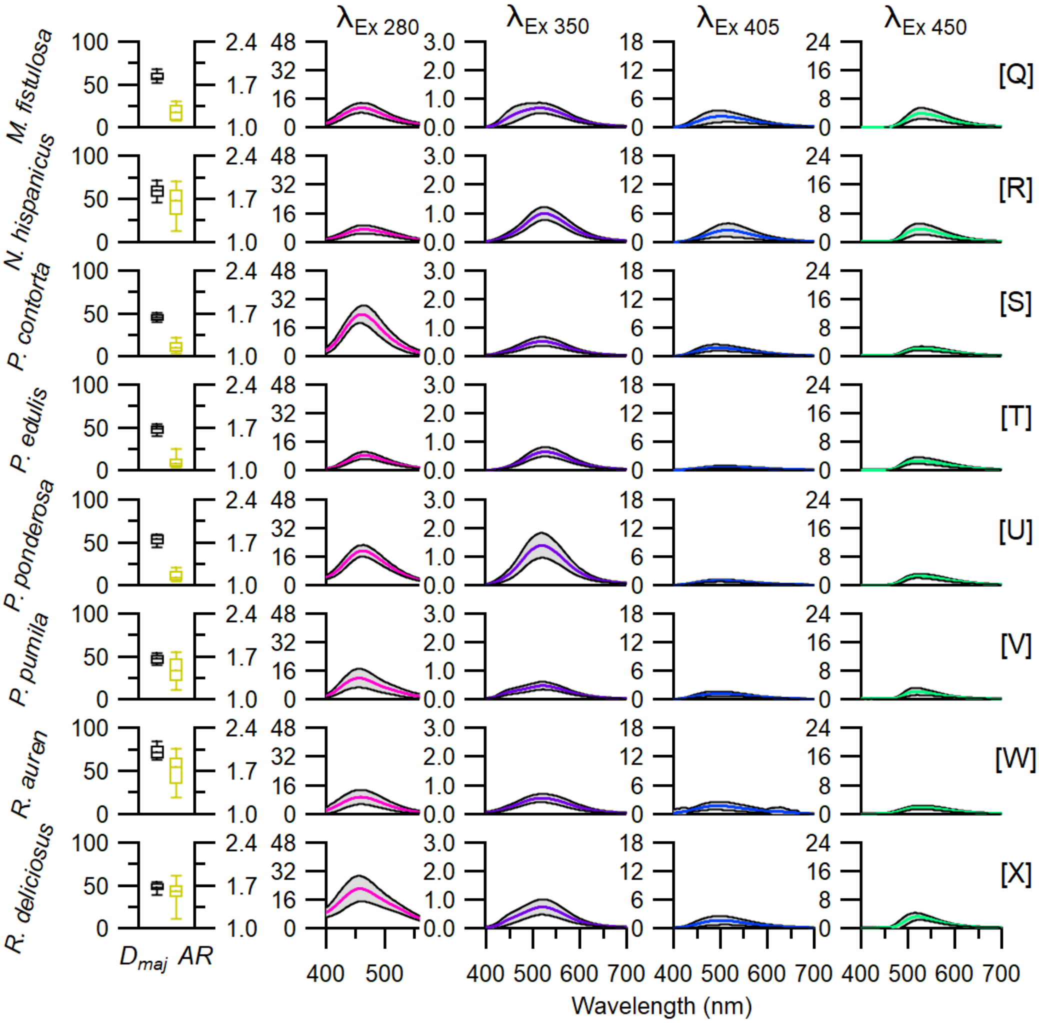
Appendix Figure B2: Comparison of pollen excitation emission modes for a typical EEM and the information available from the instrument used in this manuscript (Adapted from Pohlker et al., 2013)



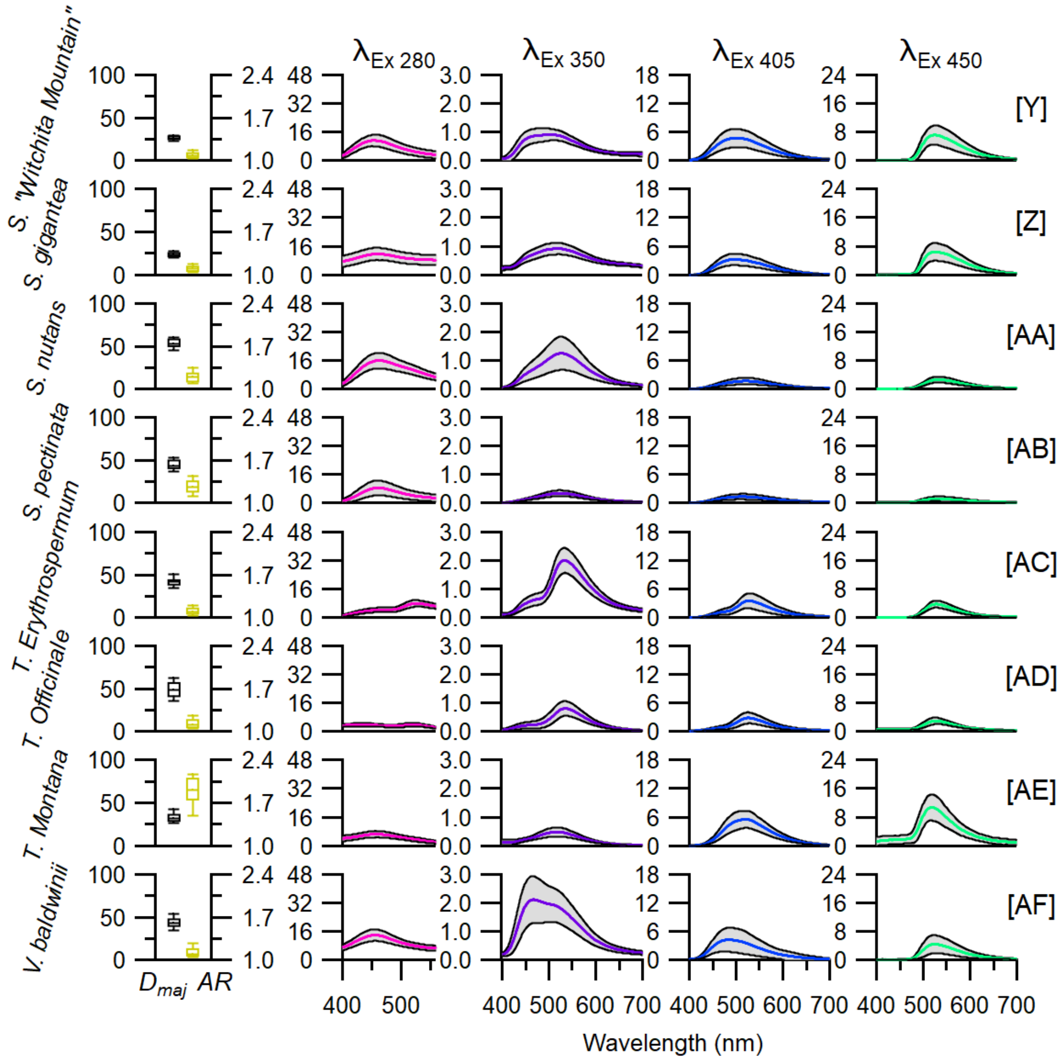
Appendix Figure B3: Species 1-8 from 2018 in the 34-species data set, with details identical to Figure 1.



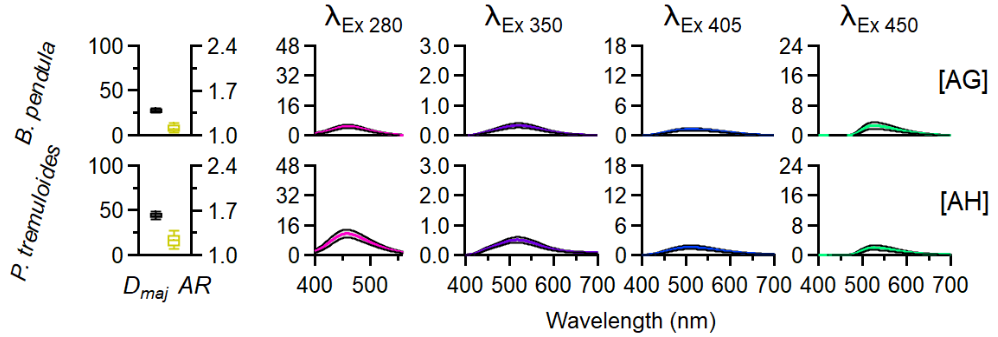
Appendix Figure B4: Species 9-16 from 2018 in the 34-species data set, with details identical to Figure 1.



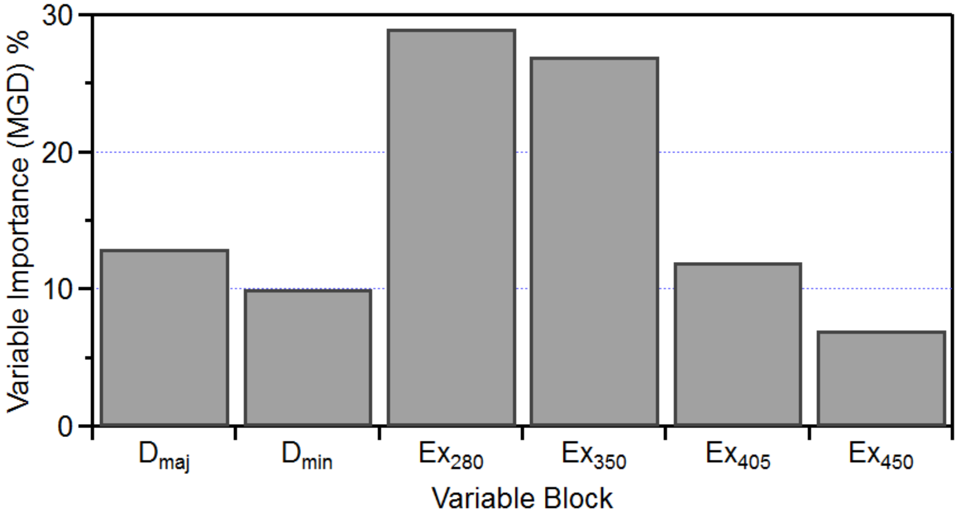
Appendix Figure B5: Species 17-24 from 2018 in the 34-species data set, with details identical to Figure 1.



Appendix Figure B6: Species 25-32 from 2018 in the 34-species data set, with details identical to Figure 1.



Appendix Figure B7: Species 31 and 32 from 2019 in the 34-species data set, with details identical to Figure 1.



Appendix Figure B8._The importance values for the sizing parameters (major and minor axis) as well as the integrated importance for the spectral emission intensity. Aspect ratio accounted for 2% of the total model but was left out for visual clarity similar to the Figure from Swanson and Huffman, 2019.

Appendix C: Towards a Compact and Automated System

C.1 Introduction

The ultimate goal of this thesis work is to lead toward the development of a smaller, inexpensive, automated pollen classification platform. The original inception and concept for this instrument was developed for usage of a common smartphone by Dr. Donald Huffman. Smartphone technology has become absolutely ubiquitous throughout the modern world. Much of this development has involved the implementation of extremely powerful camera sensors build into these smartphones. The original prototype is shown in Figure C.1a, showing the size dimensions (13.3 x 13.3 x 7.4 cm; 58 g) of the platform in reference to an iPhone 5S. Inside the box is a miniaturized version of the spectrometer described in chapter 2. A 420 nm long-pass filter present immediately under the smartphone, with a 400 nm blazed grating immediately below that. The last section of the optical components is a 10x objective lens just above the optical slide. A 650 nm red laser diode, 405 nm blue laser diode, and a small white light are present inside the box, as well as two AA batteries for the power source. Images are taken using the smartphone's camera application, and an example of an image from *Poa pratensis* particles can be seen in Figure C1b. The spectral signals from these images are seen in Figure C.1c. Considering a large number of people have smartphones with potentially powerful cameras, there is a twofold advantage here: 1) the sensor may not need inclusion in the package and 2) citizen science applications are expanded.

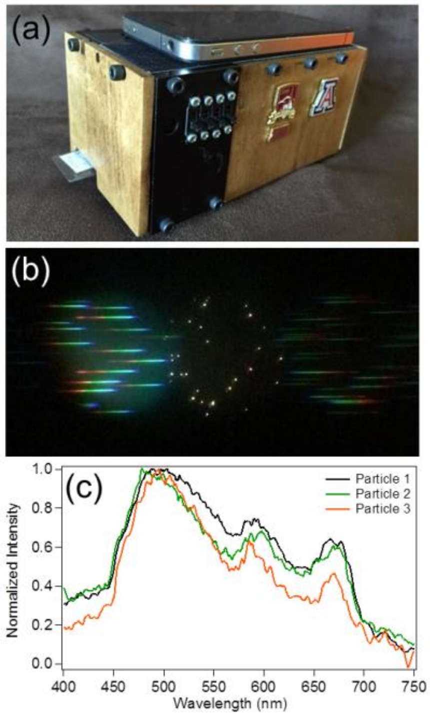


Figure C.1. The smartphone inception and initial analysis. (a) the smartphone spectrometer box developed by Dr. Donald Huffman with an iPhone 5S, (b) images produced by the spectrometer box and the iPhone 5S camera and (c) the spectra produced for three commercial *Poa prantensis* pollen particles (Reproduced from Huffman, Swanson, Huffman, 2016)

C.2 A New Sensor Using Raspberry Pi

There are obvious advantages to utilizing existing market frameworks, i.e. smartphone prevalence, in the development of instrumentation. That being said, much advancement in inexpensive computational platforms has been made in recent years. Raspberry Pi and Arduino are two examples of this, being extremely small computational platforms that have a modular interface. Development of this type of single-particle fluorescence spectrometer with a Raspberry Pi results in a loss of the citizen-science aspect of the instrumentation, but development on the collection and analysis side of the platform opens up widely.

Here, a new platform is described that is not as small or inexpensive as the miniaturized smartphone version above but allows for a much more versatile range of data collection in a way that doesn’t involve massive human input. Similar to the larger desktop version described in the main text, this instrument utilizes four excitation sources (280 nm LED; 405 nm Laser Diode; 450 nm Laser Diode; 532 nm Laser Diode). A similar grating and system of long-pass filters are also utilized to produce similar images, as well as a 10x magnification objective lens. The sensor here is a 5-megapixel arducam, with has a PiNoIR camera (colora camera; no infrared filter) and is operated directly off of a Raspberry Pi 2 system.

The overall system here was built in approximately the volume of one cubic foot and operates extremely similarly to the desktop version (no sample collection, analysis of images performed later). This can be seen in Figure C.2 showing the current setup of the instrumentation. Conceptually, there is no difference between this miniaturized version and the larger desktop research instrument in what it produces. Particles are collected on a slide prior to analysis, and then manually introduced into the system.

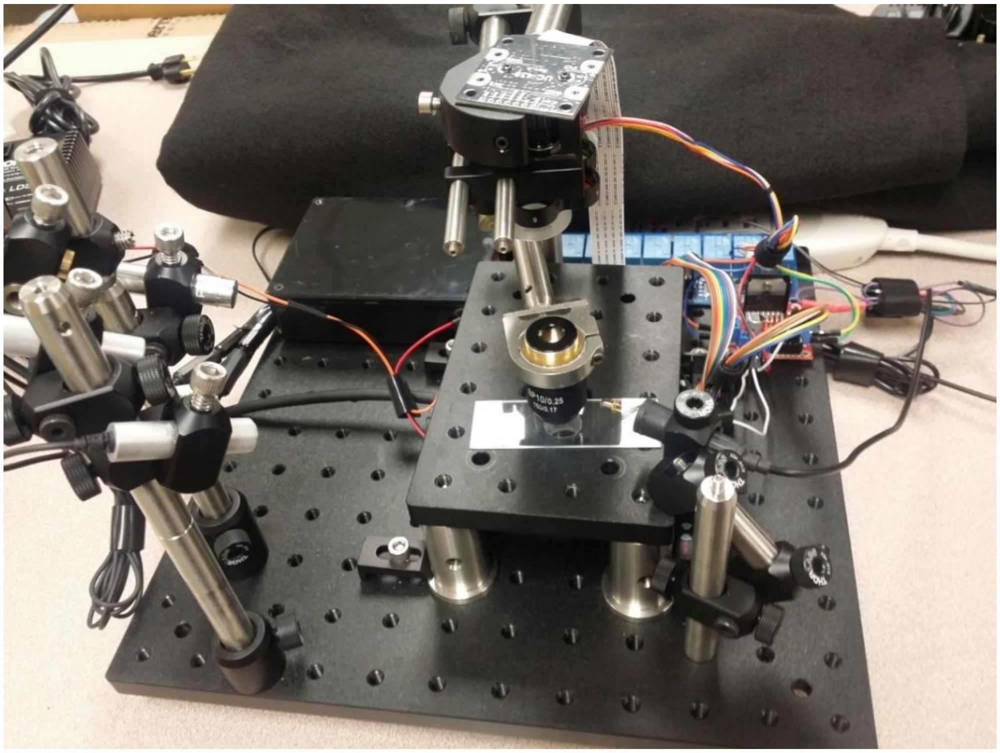


Figure C.2. Image of the current iteration of the miniaturized/automated spectrometer utilizing the Raspberry Pi framework. All laser sources are pictured on the left edge, and the 280 nm LED is pictured close to the objective lens. The Arducam can be see non the top, followed by a transmission grating, an automated filter wheel, and then some blank distance to the 10x magnification objective lens.

Appendix D: Computational Code

E.1 Open-Source R Software

The text in this appendix section represent code written within R Software, an open source statistical language platform for the R computing language. The code listed here can be pasted into R Studio and perform the clustering and classification described in the main thesis chapters.

Clustering code is printed on the following pages:

Notes regarding R Studio:

- Individual packages must be installed via install.package("[package name]") prior to use.
- Individual packages must be called on via library([package name]) when the workspaces are opened.
- Hashtags (##) denote sections of code commented out.

D.1.1 Comment Relavent packages for the following code ensemble

```
library(readr)
library(dplyr)
library(tidyr)
library(randomForest)
library(cluster)
library(factoextra)
library(party)
library(caret)
```

D.1.2 Description of code: The code provided here is for uploading and prepping/scaling data, when applicable.

```
## Loading CSV data for analysis ##
ExampleSet <- read_csv(
  "FileLocation/YourFileName.csv"
)
ExampleSet<-ExampleSet[!(is.na(ExampleSet$type)), ] ## Getting rid of Na data from original set
summary(ExampleSet) ## Summary of data

## Preparing the data for a temp file, ensuring all non-variables are listed as factors.
Temp<-0                         ## Clearing previous temp data file
Temp<-select(ExampleSet, -PartNum, -Season, -type, -allergenicity)
Temp<-as.data.frame(Temp)   ## data needs to be in DATA FRAME, not MATRIX!
Temp$type = as.character(Temp$Variety)
Temp$type = factor(Temp$Variety)
```

```
## Scaling data (when needed) for all but size
scaled<-0
scaled<-scale(select(Temp,-1,-2,-3,-4)) ## To scale size, use -1 only ##
scaled<-as.data.frame(scaled)
Temp[,5:1068]<-scaled ## re-apply  scaled data to Temp file
```

D.1.3 Description of code: The code provided here is for both the supervised and unsupervised k-means clustering methods.

```
## Set a pre-defined seed to perform clustering identically each time
set.seed(1256236)

## Use this section for supervised k-means; pre-calculated cluster centers
MeanLoc<-0
MeanLoc<-Temp%>%group_by(Variety)%>%summarise_all(mean)
MeanLoc<-data.frame(MeanLoc)

ClusterData<-0
ClusterData<-select(Temp, -Variety)

## Supervised
clust<-kmeans(ClusterData, select(MeanLoc, -Variety))   ##Perform Kmeans on (Data, select(Supervised Clusters, -Text factors))
clust$cluster         ## Show each particle's corresponding cluster.
clust$size           ## Show size of each cluster
clust$centers         ## Show center of each cluster

## Unsupervised
clust<-kmeans(ClusterData, centers=8, iter.max = 10)   ##Perform Kmeans on (Data, select(Supervised Clusters, -Text factors))
clust$cluster         ## Show each particle's corresponding cluster.
clust$size           ## Show size of each cluster
clust$centers         ## Show center of each cluster

## Check the optimal cluster number
ClusterTemp<-select(Temp, -Variety)
fviz_nbclust(ClusterTemp, cluster::pam, method = "silhouette")

## Make a confusion matrix of the data
ClusterTable<- table(clust$cluster, Temp$Variety)                    ##### Confusion Matrix
print(ClusterTable)

## White CSV files for cluster center reports and the confusion matrix
write.csv(clust$cluster, 'ClusterCenters.csv')
write.csv(ClusterTable, 'ClusterTable.csv')

## Plot cluster centers
clusplot(Temp, clust$cluster, main='2D representation of the Cluster solution',
        color=TRUE, shade=TRUE,
        labels=2, lines=0)
## Print cluster data
print(ClusterData[,1205])
```

D.1.4 Description of code: The code provided here is for the unsupervised HAC clustering utilizing Ward’s linkage.

```
library(fpc) ## load fast cluster package

set.seed(123456)

## Run HAC clustering with ward.D2 linkage.
dat.clust<-hclust(dist(Temp[,2:1068]), method = 'ward.D2')
plot(dat.clust)                          ## plot dendrogram for clustering
plot(dat.clust, cex = 0.6, hang = -1)

ClusterCut<-cutree(dat.clust, 8)  ## Cut at the desired number of clusters

ClusterTable<-table(ClusterCut, Temp$Variety) ## make confusion matrix for comparison

print(ClusterTable)
print(ClusterCut)

write.csv(ClusterTable, "ClusterTable.csv")

ClusterCut<-as.data.frame(ClusterCut)
print(dat.clust$order)

##Dendogram
require(graphics)
plot(dat.clust)
```

D.1.5 Description of code: The code provided here is for supervised gradient boostingclassification.

```
library(gbm)
library(caret)
library(doParallel)

## for segmenting commercial pollen train:test sets ##
train <- Temp[1:134,] #### 66/33
test <- Temp[135:204,]

train <- Temp[1:153,] #### 75/25
test <- Temp[154:204,]

## Gradient boosting by species ##
mod_gb <- gbm(Variety~.,
            data = train,
            distribution = "multinomial",
            interaction.depth = 1,
            cv.folds = 0,
            shrinkage = .001,
            n.minobsinnode = 10,
            n.trees = 500)

## make the ideal tree iteration a variable
best.gbm.predict<-gbm.perf(mod_gb, method = "cv")

## show statistics on the models
print(mod_gb)
t <- pretty.gbm.tree(mod_gb, i=25)

## Show the exponential loss curve
print(mod_gb$train.error)
sqrt(min(mod_gb$cv.error))
gbm.perf(mod_gb, method = "cv")
rowMax<-apply(Predictions, 1, max)
rowMax<-as.data.frame(rowMax)

## predict new data to the gradient boosting model based on the ideal iteration
Predictions <- predict.gbm(object = mod_gb,
                   newdata = test,
                   n.trees = best.gbm.predict,
                    type = "response")

## Prep data for confusion matrix
result <- cbind(train[,1], Predictions)
p.Predictions <- apply(Predictions, 1, which.max)
p.Predictions<-as.data.frame(p.Predictions)
result <- cbind(train[,1], p.Predictions)
write.csv(result, "GradientBoostResult.csv")

## Confusion matrix and print
ClusterTable<-table(result)
print(ClusterTable)

## Develop and save CSV version of variable importance file
print(mod_gb)
sqrt(min(mod_gb$cv.error))
gbm.perf(mod_gb, method = "cv")
summary(mod_gb)
VariableImp<-summary(mod_gb)
write.csv(VariableImp, "ImportanceFile.csv")
```

D.1.6 Description of code: The code provided here is for supervised random forest classification.

```
## For raw data
train <- Temp
test <- Temp

## for 75/25 train test
train <- Temp[1:153,]
test <- Temp[154:204,]

## Clear any previous model
output.cforest<-0


## Develop new random forest model based on training data
## 1000 tree forest with 5 variables examined at any one time
output.cforest<-cforest(Variety ~ .,
                  data = train,
                  controls=cforest_unbiased(ntree=1000,mtry=5))

## Find variable importance within model
variables<-varimp(output.cforest)
print(variables)
write.csv(variables, "ImportanceFile.csv")

####### Shows the Error for Training Set #######
predictions<-predict(output.cforest, newdata = train, type = "response", OOB = TRUE)
predictions<-as.data.frame(predictions)
varieties<-train$Variety  #### edit this for diff types! (“train$’type’”)
varieties<-as.data.frame(varieties)
ClusterTable<-cbind(varieties,predictions)
ClusterTable<-as.data.frame(ClusterTable)
write.csv(ClusterTable, "TrainingSet.csv")
print(ClusterTable)
ClusterTable<-table(ClusterTable)

## Create confusion matrix for training set
result <- cbind(train[,1], predictions)
```

```
print(result)
p.train<-(train)
p.Predictions <- apply(predictions, 1, which.max)
p.Predictions<-as.data.frame(p.Predictions)
result <- cbind(train[,1], p.Predictions)
ClusterTable<-table(result)
write.csv(Result, "ClusterResults.csv")
print(ClusterTable)
write.csv(ClusterTable, "ConfusionMatrix.csv")

####### Predict a test, or new data set, to the model #######
predictions<-0
predictions<-predict(output.cforest, newdata = test, type = "response", OOB = TRUE)
predictions<-as.data.frame(predictions)
varieties<-test$Variety
varieties<-as.data.frame(varieties)
ClusterTable<-cbind(varieties,predictions)
ClusterTable<-as.data.frame(ClusterTable)
write.csv(ClusterTable, "TestSet.csv")
print(ClusterTable)
```

D.2 Igor Pro Software Tools

D.2.1 Description of Functions: Code intended to make waves compliant for the baseline subtraction code.

```
Function SmoothWaves()
        Variable n
        String NameStrY, NewSizeWave, NameStrX, NewSizeWaveX
        Wave BluePts

        For (n=1;n<=DimSize(BluePts,0);n+=1)
                NameStrY = "p" + num2str(n) + "y450_final"                // Create a string out of p[n]yCalib, n being
the particle number
                Duplicate/O $NameStrY, $NameStrY+"_smth" ;DelayUpdate
                Smooth 100, $NameStrY+"_smth"

                NameStrY = "p" + num2str(n) + "y405_final"
                Duplicate/O $NameStrY, $NameStrY+"_smth" ;DelayUpdate
                Smooth 500, $NameStrY+"_smth"

                NameStrY = "p" + num2str(n) + "y350_final"
                Duplicate/O $NameStrY, $NameStrY+"_smth" ;DelayUpdate
                Smooth 500, $NameStrY+"_smth"

                NameStrY = "p" + num2str(n) + "y280_final"
                Duplicate/O $NameStrY, $NameStrY+"_smth" ;DelayUpdate
                Smooth 100, $NameStrY+"_smth"
        EndFor
End

Function DisplaySmth()
        Variable n
        String NameStrY, NewSizeWave, NameStrX, NewSizeWaveX
        Wave BluePts

        Display

        For (n=1;n<=DimSize(BluePts,0);n+=1)
                NameStrY = "p" + num2str(n) + "y280_final_smth"
                AppendToGraph $NameStrY
        EndFor

        Display

        For (n=1;n<=DimSize(BluePts,0);n+=1)
                NameStrY = "p" + num2str(n) + "y350_final_smth"
                AppendToGraph $NameStrY
        EndFor

        Display

        For (n=1;n<=DimSize(BluePts,0);n+=1)
                NameStrY = "p" + num2str(n) + "y405_final_smth"
                AppendToGraph $NameStrY
        EndFor

        Display

        For (n=1;n<=DimSize(BluePts,0);n+=1)
                NameStrY = "p" + num2str(n) + "y450_final_smth"
                AppendToGraph $NameStrY
```

```
		EndFor
End


Function PasteFinal()
		Variable n
		String NameStrY
		Wave Bluepts

		Edit

		For (n=1; n<=DimSize(BluePts,0);n+=1)
				NameStrY = "p" + num2str(n) + "y280_final_smth_sub"
				AppendToTable $NameStrY
		Endfor
		edit
		For (n=1; n<=DimSize(BluePts,0);n+=1)
				NameStrY = "p" + num2str(n) + "y350_final_smth_sub"
				AppendToTable $NameStrY
		Endfor
		edit
		For (n=1; n<=DimSize(BluePts,0);n+=1)
				NameStrY = "p" + num2str(n) + "y405_final_smth_sub"
				AppendToTable $NameStrY
		Endfor
		edit
		For (n=1; n<=DimSize(BluePts,0);n+=1)
				NameStrY = "p" + num2str(n) + "y450_final_smth_sub"
				AppendToTable $NameStrY
		Endfor

End
```

D.2.3 Description of Functions: Code to import .tiff files for each individual emission wave and calibration waves, to fully calibrate all emission spectra based off those calibration images, and to fully measure the size of the particles and calibrate size and source power density.

```
//******************************************************************************
// FULL FUNCTION TO CALIBRATION
//******************************************************************************


Function CropAnalysis(pathName, nameWaveAfterFileName, displayImages) // ("", 1, 0)
 // IMPORT TIFFS FROM SUBFOLDER
		String pathName
		Variable nameWaveAfterFileName, displayImages

		if (strlen(pathName) == 0 )
								// see if there's a given path/folder.  Quotations ("") returns
				NewPath/O/M="Choose a folder containing TIFFs" LoadInexedTIFFPath			// Prompt
diag box asking for folder input
				If (V_Flag != 0)
						Return -1
				EndIf
				pathName = "LoadInexedTIFFPath"
								// Make PathName the folder selected
		EndIf

		Variable Count
```

```
        Count = 1

        String fileName, list, fileLoc
        Variable index
        Wave FileSort

        Make/O/N=1000 Exposure280                       // Make a wave for exposure time for 280
        Make/O/N=1000 Exposure350                       // 350
        Make/O/N=1000 Exposure405                       // 405
        Make/O/N=1000 Exposure450                       // 450

        Exposure280 = NaN                                       // Make them all NaN
        Exposure350 = NaN
        Exposure405 = NaN
        Exposure450 = NaN

        index = -1

        fileLoc = IndexedFile($pathName, -1, ".tif")
        fileLoc = SortList(FileLoc, ";", 16)
        wave FileSort = $fileLoc

        Variable n
        n = 0

        do
                fileName = StringFromList(n, FileLoc, ";")              // find ".tif" files in the specified folder
                if (strlen(fileName) == 0)                                                                      // If a ".tif"
image exists, keep going, if not, break the loop
                        break
        // break the loop
                endif

                ImageLoad/P=$pathName/T=TIFF/N=image fileName  // Load the image into igor, specifically Tiffs

                if (V_Flag > 0)
                        string name = StringFromList(0, S_waveNames)            //
                        wave w = $name

                        if (nameWaveAfterFileName)
                                string desiredName = S_fileName
                        // desiredName is equal to the full image name + ".tif"
                                String FileNameStr = S_fileName
                        // FileNameSt = file name + ".tif"
                                desiredName = ParseFilepath(3, S_filename, ":", 0, 0)
        // Parse the file's name and put it into S_Filename

                                                String NoTiff

                                                NoTiff = RemoveEnding(S_Filename, ".tif")
        // Remove the ".tif" ending on each loaded image

                                                String expr= "%s %e"
                                // Make expr a string EXPRESSION equal to "%s %e" // %s stores all text to the next
white space, %e catches the number exp behind it
                                                String LaserExSt
                                                Variable ExpTime

                                                sscanf NoTiff, Expr, LaserExSt, ExpTime
        // Separate the string NoTiff by Expr (%s %e) and split it into LaserExSt string and ExpTime string
```

```
                        String ExcValue280 = num2str(count+1) + "y280"
    // Make strings for each wave type per particle
                        String Excvalue350 = num2str(count+1) + "y350"
                        String Excvalue405 = num2str(count+1) + "y405"
                        String Excvalue450 = num2str(count+1) + "y450"
                        String ExcValueCalib = num2Str(count+1) + "yCalib"


                        If (Stringmatch(LaserExSt, ExcValue280))
    // Add the exposure time number to each exposure time wave.
                            Exposure280[Count] = ExpTime
                        ElseIf (Stringmatch(LaserExSt, ExcValue350))
                            Exposure350[Count] = ExpTime
                        ElseIf (Stringmatch(LaserExSt, ExcValue405))
                            Exposure405[Count] = ExpTime
                        Elseif (Stringmatch(LaserExSt, ExcValue450))
                            Exposure450[Count] = ExpTime
                            Count +=1
                        ElseIf (StringMatch(LaserExSt, ExcValueCalib))
                        EndIf

                        //String expr="([[:alnum:]]+) ([[:digit:]]+)"
                        //String LaserExSt, ExpTime, GainValue, Tiff
                        //SplitString/E=expr NoTiff, LaserExSt, ExpTime

                        if (Exists(LaserExSt) != 0)
                // If the desired name exists, this wave gets an identical name with n1, n2, n... starting
from 0
                            LaserExSt = UniqueName(LaserExSt, 1, 0)
                        EndIf

                        Rename w, $LaserExSt


                EndIf

            if (displayImages)
                NewImage w
            endif
        endif
        n += 1

    while (1)

    Variable NaNdel
    NaNdel = 1

    For (NaNdel = 1 ; NaNdel <= numpnts(Exposure280) ; NaNdel+= 1)                    // Go from
1 to the length of exposure280
        If (numtype(Exposure280[NaNdel-1]) == 2)
    // Check for NaNs
            Deletepoints NaNdel-1,1,Exposure280
        // Delete NaNs
                NaNdel-=1
        Endif
    Endfor

    NaNdel = 1

    For (NaNdel = 1 ; NaNdel <= numpnts(Exposure350) ; NaNdel+= 1)
        If (numtype(Exposure350[NaNdel-1]) == 2)
```

```
                        Deletepoints NaNdel-1,1,Exposure350
                                NaNdel-=1
                Endif
        Endfor

        NaNdel = 1

        For (NaNdel = 1 ; NaNdel <= numpnts(Exposure405) ; NaNdel+= 1)
                If (numtype(Exposure405[NaNdel-1]) == 2)
                        Deletepoints NaNdel-1,1,Exposure405
                                NaNdel-=1
                Endif
        Endfor

        NaNdel = 1

        For (NaNdel = 1 ; NaNdel <= numpnts(Exposure450) ; NaNdel+= 1)
                If (numtype(Exposure450[NaNdel-1]) == 2)
                        Deletepoints NaNdel-1,1,Exposure450
                                NaNdel-=1
                Endif
        Endfor

        // Calibrating Factors fix
        Exposure280 *=1000                      // Change seconds to milliseconds
        Exposure350 *=1000
        Exposure405 *=1000
        Exposure450 *=1000

        Make/O/N=4 PowerDensity
        PowerDensity[0] = 0.004                 // PD factor correction for 280
        PowerDensity[1] = 0.5                           // PD factor correction for 350
        PowerDensity[2] = 4.5                           // PD factor correction for 405
        PowerDensity[3] = 0.3                           // PD factor correction for 450

        Variable rn, t, r, y, j, b, o, h
        String wCalib, Calib, w450, w405, w350, w280, yCalib, y450, y405, y350, y280, pXRef, CurrentRef

        b = 0
        j = 1
        o = 1
        n = 1

        For (o=1; o <= (count+1) ; o +=1)               // Analyze profile and rename all spectral swath crops to waves with
p(n)yXXX and p(n)xXXX individual waves

                yCalib = Num2Str(o)+"yCalib"
                y450 = Num2Str(o)+"y450"
                y405 = Num2Str(o)+"y405"
                y350 = Num2Str(o)+"y350"
                y280 = Num2Str(o)+"y280"

                J = WaveExists($yCalib)

                If (j == 1)

                Wave iCalib = $yCalib
                Wave i450 = $y450
                Wave i405 = $y405
                Wave i350 = $y350
                Wave i280 = $y280
        If  (j == 1)
                h= DimSize(iCalib,0)

                Make/O /N=(h) XRef
// Make XRef wave equal to r (XRef is X dimension for spectra)
                t = 1
        Endif

                        Variable Refn

                        For (Refn=1;(Refn-1)<DimSize($yCalib, 0);Refn+=1)
    // Start at 1 and make the xRef value in pixels
                            XRef[Refn-1]=t

                                t+=1
                        Endfor

                CurrentRef = "p"+Num2Str(o)+"xRef"
                Make/O /N=(h) $CurrentRef = xRef

                wCalib = "p"+Num2Str(o)+"yCalib"
                w450 = "p"+Num2Str(o)+"y450"
                w405 = "p"+Num2Str(o)+"y405"
                w350 = "p"+Num2Str(o)+"y350"
                w280 = "p"+Num2Str(o)+"y280"

                MatrixOP/O AveragesWave = sumRows(iCalib)/numCols(iCalib)
                Make/O/N=(DimSize($CurrentRef,0)) $wCalib = AveragesWave

                MatrixOP/O AveragesWave = sumRows(i450)/numCols(i450)
                Make/O/N=(DimSize($CurrentRef,0)) $w450 = AveragesWave

                MatrixOP/O AveragesWave = sumRows(i405)/numCols(i405)
                Make/O/N=(DimSize($CurrentRef,0)) $w405 = AveragesWave

                MatrixOP/O AveragesWave = sumRows(i350)/numCols(i350)
                Make/O/N=(DimSize($CurrentRef,0)) $w350 = AveragesWave

                MatrixOP/O AveragesWave = sumRows(i280)/numCols(i280)
                Make/O/N=(DimSize($CurrentRef,0)) $w280 = AveragesWave

                b += 1
        EndIf
EndFor

Make/o/n=1000 SizeWave
SizeWave = NaN

Wave Fit_CurveFitting
String SizeMeas405
Variable l
n = 1
j = 1

Variable Check
Check = 1

Variable CountingUp, DelPts
Wave MajMom, MinMom, AngMom, AspectRatio, YSize, Degrees, XSize
```

```
            CountingUp = 1
        Make/O/N=50 MajMom, MinMom, AngMom, AspectRatio, YSize, Degrees, XSize, BlueCalib, RedCalib

        Degrees = NaN
        MajMom = NaN
        MinMom = NaN
        AngMom = NaN
        AspectRatio = NaN
            YSize = NaN
            XSIze = NaN
            BlueCalib = NaN
            RedCalib = NaN

       do
          String Measured, BinMeas
         Measured = num2str(CountingUp) + "yCalib"

          Wave w0=$Measured
           if(WaveExists(w0)==0)
              break
           endif

            //////////////////////////////////////////////////////////////
         ImageThreshold/Q/T=25/I w0                                                  // EDIT THIS TO CHANGE
THRESHOLD //
         Wave M_ImageThresh                    //////////////////////////////////////////////////////////////
         Duplicate/O M_ImageThresh, RedPointWave

         print "THIS RIGHT HERE" + num2str(CountingUp)

         ImageThreshold/Q/T=25/I w0
         Duplicate/O M_ImageThresh, BluePointWave

         Redimension/N=(-1,-1) RedPointWave
         Redimension/N=(-1,-1) BluePointWave


         Duplicate/O/R=(400,600)(0,104) RedPointWave, ParticleMeasure // (X)(Y) points
         Duplicate/O/R=(400,600)(0,104) RedPointWave, RedMeasure
         Duplicate/O/R=(0,200)(0,104) BluePointWave, BlueMeasure

         ImageAnalyzeParticles /E/W/Q/M=3/A=100 stats, BlueMeasure
            Wave M_Moments
            BlueCalib[CountingUp-1] = M_Moments[0]
         ImageAnalyzeParticles /E/W/Q/M=3/A=100 stats, RedMeasure
            RedCalib[CountingUp-1] = M_Moments[0]+400
         ImageAnalyzeParticles /E/W/Q/M=3/A=100 stats, ParticleMeasure

           MajMom[CountingUp-1] = M_Moments[2]*2
           MinMom[CountingUp-1] = M_Moments[3]*2

           If (M_Moments[4]>pi)
                    M_Moments[4] = M_Moments[4]/pi
                    AngMom[CountingUp-1] = M_Moments[4]

              Else
                    AngMom[CountingUp-1] = M_Moments[4]

           EndIf
```

```
             AspectRatio[CountingUp-1] = MajMom[CountingUp-1]/MinMom[CountingUp-1]

              Print CountingUp

          Wave W_YMax, W_YMin, W_XMax, W_XMin
          YSize[CountingUp-1] = W_YMax[0]-W_YMin[0]
          XSize[CountingUp-1] = W_XMax[0]-W_XMin[0]

                    Degrees[CountingUp-1] = AngMom[CountingUp-1]*(180/pi)

    CountingUp +=1
          while(1)

  DelPts = 1

          For (DelPts = 1 ; DelPts <= numpnts(MajMom) ; DelPts+= 1)                                      // Go from 1 to the
length of exposure280
                    If (numtype(MajMom[DelPts-1]) == 2)
          // Check for NaNs
                              DeletePoints DelPts-1,1,MajMom
                    // Delete NaNs
                                        DelPts-=1
                    Endif
          Endfor

  For (DelPts = 1 ; DelPts <= numpnts(BlueCalib) ; DelPts+= 1)                                  // Go from 1 to the length of
exposure280
                    If (numtype(BlueCalib[DelPts-1]) == 2)
          // Check for NaNs
                              DeletePoints DelPts-1,1,BlueCalib
                    // Delete NaNs
                                        DelPts-=1
                    Endif
          Endfor

  For (DelPts = 1 ; DelPts <= numpnts(RedCalib) ; DelPts+= 1)                                   // Go from 1 to the length of
exposure280
                    If (numtype(RedCalib[DelPts-1]) == 2)
          // Check for NaNs
                              DeletePoints DelPts-1,1,RedCalib
                    // Delete NaNs
                                        DelPts-=1
                    Endif
          Endfor

          For (DelPts = 1 ; DelPts <= numpnts(MinMom) ; DelPts+= 1)                                      // Go from 1 to the
length of exposure280
                    If (numtype(MinMom[DelPts-1]) == 2)
          // Check for NaNs
                              DeletePoints DelPts-1,1,MinMom
                    // Delete NaNs
                                        DelPts-=1
                    Endif
          Endfor

          For (DelPts = 1 ; DelPts <= numpnts(AngMom) ; DelPts+= 1)                                      // Go from 1 to the
length of exposure280
                    If (numtype(AngMom[DelPts-1]) == 2)
          // Check for NaNs
                              DeletePoints DelPts-1,1,AngMom
                    // Delete NaNs
```

```
                DelPts-=1
            Endif
        Endfor

        For (DelPts = 1 ; DelPts <= numpnts(AspectRatio) ; DelPts+= 1)                         // Go from 1 to the
    length of exposure280
            If (numtype(AspectRatio[DelPts-1]) == 2)
        // Check for NaNs
                DeletePoints DelPts-1,1,AspectRatio
            // Delete NaNs
                    DelPts-=1
            Endif
        Endfor

        For (DelPts = 1 ; DelPts <= numpnts(YSize) ; DelPts+= 1)                               // Go from 1 to the
    length of exposure280
            If (numtype(YSize[DelPts-1]) == 2)
        // Check for NaNs
                DeletePoints DelPts-1,1,YSize
        // Delete NaNs
                    DelPts-=1
            Endif
        Endfor

        For (DelPts = 1 ; DelPts <= numpnts(XSize) ; DelPts+= 1)                               // Go from 1 to the
    length of exposure280
            If (numtype(XSize[DelPts-1]) == 2)
        // Check for NaNs
                DeletePoints DelPts-1,1,XSize
        // Delete NaNs
                    DelPts-=1
            Endif
        Endfor

        For (DelPts = 1 ; DelPts <= numpnts(Degrees) ; DelPts+= 1)                             // Go from 1 to the
    length of exposure280
            If (numtype(Degrees[DelPts-1]) == 2)
        // Check for NaNs
                DeletePoints DelPts-1,1,Degrees
        // Delete NaNs
                    DelPts-=1
            Endif
        Endfor

        For (n=1;n<=(count+1); n+=1)

            Make/O/N=2 DestWave
            Variable Mid

            Sizemeas405 = num2str(n)+"y405"
            Wave MeasuredWave = $SizeMeas405
            String pSize, pSizeRef, SizeMeas2

            pSize = num2str(n)+"pSize"

            MatrixOP/O sizingwave = sumCols(MeasuredWave)/numRows(MeasuredWave)
            Make/O/N=(DimSize($Sizemeas405,1)) $pSize = SizingWave

            Wave CurveFitting = $pSize

            CurveFit/NTHR=0 gauss CurveFitting /D
```

```
                String SizeMeas = "fit_" + pSize
                Wave Subtractor = $SizeMeas
                Wavestats Subtractor                                                    // Measure the Guassians'
characteristics

                CurveFitting -= V_Min

                CurveFit/NTHR=0 gauss CurveFitting /D
                SizeMeas2 = "fit_" + pSize
                Wave Subtractor = $SizeMeas2
                Wavestats Subtractor

                Mid = V_max/2
                FindLevels/D=destWave $SizeMeas, Mid

                SizeWave[n-1] = (DestWave[1]-DestWave[0])

                Check =+1

        EndFor

        NanDel = 1

        For (NaNdel = 1 ; NaNdel <= numpnts(SizeWave) ; NaNdel+= 1)
                If (numtype(SizeWave[NaNdel-1]) == 2)
                        Deletepoints NaNdel-1,1,SizeWave
                                NaNdel-=1
                Endif
        Endfor

        String NameStr, p, Fit, Calibrate

        j = 1
        Make /O /N=1000 BluePts
        BluePts = NaN
        Make /O /N=1000 RedPts
        RedPts = NaN

        n = 1
        For (n = 1; n<=Count ; n += 1)                                  // Still need to try to recognize if the wave
exists.

                NameStr = "p" + num2str(n) + "yCalib"                   // Create a string out of p[n]yCalib, n being the particle
number
                Wave Calibrating = $NameStr                             // Make the NameStr into a reference and
make the Wave Calibrating that string

                j = WaveExists(Calibrating)


                if (j ==1)
                        CurveFit/NTHR=0 gauss  Calibrating[0,200] /D // Measure the Gaussian around the blue calib point
                        Calibrate = "fit_" + NameStr
                        Wave CalibWave = $Calibrate
                        Wavestats CalibWave                                     // Measure the
Guassians' characteristics
                        BluePts[n-1] = V_MaxLoc                                 // Add the Max
Location Calib point to the BluePts wave

                        CurveFit/NTHR=0 gauss  Calibrating[400,650]  /D // Measure the Gaussian around the red calib
point
```

```
                Calib = "fit_" + NameStr
                Wave CalibWave = $Calib
                Wavestats CalibWave
                RedPts[n-1] = V_MaxLoc
        EndIf
EndFor

NaNdel =1

For (NaNdel = 1 ; NaNdel <= numpnts(Bluepts) ; NaNdel+= 1)
        If (numtype(Bluepts[NaNdel-1]) == 2)
                Deletepoints NaNdel-1,1,Bluepts
                        NaNdel-=1
        Endif
Endfor

NaNdel =1

For (NaNdel = 1 ; NaNdel <= numpnts(Redpts) ; NaNdel+= 1)
        If (numtype(Redpts[NaNdel-1]) == 2)
                Deletepoints NaNdel-1,1,Redpts
                        NaNdel-=1
        Endif
Endfor

                // For deleting NaN poits.  Not needed here now.
                //********************************************************
                //      For (n = 1 ; n <= numpnts(BluePts) ; n+= 1)
                //              If (numtype(BluePts[n-1]) == 2)
                //                      Deletepoints n-1,1,BluePts
                //                      n-=1
                //              Endif
                //      Endfor
                //
                //      For (n = 1 ; n <= numpnts(RedPts) ; n+= 1)
                //              If (numtype(Redpts[n-1]) == 2)
                //                      Deletepoints n-1,1,RedPts
                //                      n-=1
                //              Endif
                //      Endfor
                //********************************************************
Variable c, e                                           // n, c, r
Wave BluePts, RedPts
String NameStr450, NameStr405, NameStr350, NameStr280
.8-403.457

y = 0
For (y = 0; y <= DimSize(BluePts,0); y += 1)

                NameStr450 = "p" + num2str(y) + "xREF"

                Wave CalibRef = $NameStr450

                e = (RedCalib[y]-BlueCalib[y])

                CalibRef -=BlueCalib[y]

                CalibRef *= (c/e)

                CalibRef +=403.457
```

```
EndFor


String NameStrX, NameStrY, NewSizeWaveX, NewSizeF
Wave BluePts, YSize, XSize, FSize

//////////////////////////////////////////////////
// Copy command for new sizing //
//////////////////////////////////////////////////
String NewSizeWave, NameStrXSize, NameStrFSize

For (n = 1; n <= DimSize(BluePts, 0); n += 1)

        NameStrY = "p" + num2str(n) + "y450"
        NewSizeWave = "p" + num2str(n) + "y450_New"
        NameStrXSize = "p" + num2str(n) + "y450_NewX"
        NameStrFSize = "p" + num2str(n) + "y450_NewF"

        Duplicate/O $NameStrY, $NewSizeWave
        Duplicate/O $NameStrY, $NameStrXSize
        Duplicate/O $NameStrY, $NameStrFSize

EndFor

For (n = 1; n <= DimSize(BluePts, 0); n += 1)

        NameStrY = "p" + num2str(n) + "y405"
        NewSizeWave = "p" + num2str(n) + "y405_New"

        NameStrXSize = "p" + num2str(n) + "y405_NewX"
        NameStrFSize = "p" + num2str(n) + "y405_NewF"

        Duplicate/O $NameStrY, $NewSizeWave
        Duplicate/O $NameStrY, $NameStrXSize
        Duplicate/O $NameStrY, $NameStrFSize

EndFor

For (n = 1; n <= DimSize(BluePts, 0); n += 1)

        NameStrY = "p" + num2str(n) + "y350"
        NewSizeWave = "p" + num2str(n) + "y350_New"

        NameStrXSize = "p" + num2str(n) + "y350_NewX"
        NameStrFSize = "p" + num2str(n) + "y350_NewF"

        Duplicate/O $NameStrY, $NewSizeWave
        Duplicate/O $NameStrY, $NameStrXSize
        Duplicate/O $NameStrY, $NameStrFSize

EndFor

For (n = 1; n <= DimSize(BluePts, 0); n += 1)

        NameStrY = "p" + num2str(n) + "y280"
        NewSizeWave = "p" + num2str(n) + "y280_New"
        NameStrXSize = "p" + num2str(n) + "y280_NewX"
        NameStrFSize = "p" + num2str(n) + "y280_NewF"

        Duplicate/O $NameStrY, $NewSizeWave
        Duplicate/O $NameStrY, $NameStrXSize
        Duplicate/O $NameStrY, $NameStrFSize

EndFor

String PeakSW, PeakY, PeakX, PeakXRef, PeakF
Wave Bluepts, AngMom

Make/O/N=(DimSize(AngMom, 0)) PeaksSW, PeaksY, PeaksX, PeaksF

PeaksSW = NaN
PeaksY = NaN
PeaksX = NaN
PeaksF = NaN
Variable up
Up = 0

Wave Degrees, YSize, XSize, FSize

        Up = 0

        Make/O FactorSplitY, FactorSplitX, FSize

        FactorSplitY = NaN
        FactorSplitX = NaN
        FSize = NaN

For     (up = 0; up<DimSize(Degrees, 0); up +=1)
        If (Degrees[Up]<90)
                FactorSplitY[up] = Degrees[up]/90
                FactorSplitX[up]  = 1-FactorsplitY[up]
        ElseIf (Degrees[Up]>90)
                FactorSplitX[up] = Degrees[up]/180
                FactorSplitY[up]  = 1-FactorsplitX[up]
        EndIf

        FSize[up] = ((FactorSplitY[up]*YSize[up]))+((FactorSplitX[up]*XSize[up]))
EndFor


For (NaNdel = 1 ; NaNdel <= numpnts(FactorSplitX) ; NaNdel+= 1)
        If (numtype(FactorSplitX[NaNdel-1]) == 2)
                Deletepoints NaNdel-1,1,FactorSplitX
                        NaNdel-=1
        Endif
Endfor

For (NaNdel = 1 ; NaNdel <= numpnts(FactorSplitY) ; NaNdel+= 1)
        If (numtype(FactorSplitY[NaNdel-1]) == 2)
                Deletepoints NaNdel-1,1,FactorSplitY
                        NaNdel-=1
        Endif
Endfor

For (NaNdel = 1 ; NaNdel <= numpnts(FSize) ; NaNdel+= 1)
        If (numtype(FSize[NaNdel-1]) == 2)
                Deletepoints NaNdel-1,1,FSize
                        NaNdel-=1
        Endif
Endfor

////////////////////////////////////////////////
```

```
        // Display & Calibration Section //
        ////////////////////////////////////////////////

        Display
        j = 1
        n = 1
        For (n = 1; n <=DimSize(BluePts,0); n += 1)                                  // Still need to try to recognize if
the wave exists.

                        NameStrY = "p" + num2str(n) + "y450"               // Create a string out of p[n]yCalib, n being
the particle number
                        NewSizeWave = "p" + num2str(n) + "y450_New"
                        Wave DisplayNew = $NewSizeWave
                        NewSizeWaveX = "p" + num2str(n) + "y450_NewX"
                        NameStrX = "p" + num2str(n) + "xRef"
                        Wave DisplayY = $NameStrY
                        Wave DisplayX = $NameStrX
                        Wave DIsplaynewx = $NewSizeWaveX

                        NewSizeF = "p" + num2str(n) + "y450_NewF"
                        Wave DisplayNewF = $NewSizeF

                        j = (waveexists($NameStrY))

                        if (j != 1)
                                Break
                        EndIf

                        //Quick baseline subtraction
                        Wavestats DisplayY
                        DisplayY -=V_min
                        DisplayY /=PowerDensity[3]
                        DisplayY *=100000

                        DisplayY /= SizeWave[n-1]
                        DisplayY /= Exposure450[n-1]

                        Wavestats DisplayNew
                        DisplayNew -=V_min
                        DisplayNew /=PowerDensity[3]
                        DisplayNew *=100000

                        DisplayNew /= YSize[n-1]
                        DisplayNew /= Exposure450[n-1]

                        Wavestats DisplayNewX
                        DisplayNewx -=V_min
                        DisplayNewx /=PowerDensity[3]
                        DisplayNewx *=100000

                        DisplayNewx /= XSize[n-1]
                        DisplayNewx /= Exposure450[n-1]

                        Wavestats DisplayNewF
                        DisplayNewF -=V_min
                        DisplayNewF /=PowerDensity[3]
                        DisplayNewF *=100000

                        DisplayNewF /= FSize[n-1]
                        DisplayNewF /= Exposure450[n-1]
```

```
            AppendToGraph DisplayY vs DisplayX
    EndFor

    Display
    n = 1
    For (n = 1; n <=DimSize(BluePts,0); n += 1)                    // Still need to try to
recognize if the wave exists.

            NameStrY = "p" + num2str(n) + "y405"            // Create a string out of p[n]yCalib, n being
the particle number
            NewSizeWave = "p" + num2str(n) + "y405_New"
            Wave DisplayNew = $NewSizeWave
            NewSizeWaveX = "p" + num2str(n) + "y405_NewX"
            NameStrX = "p" + num2str(n) + "xRef"
            Wave DisplayY = $NameStrY
            Wave DisplayX = $NameStrX
            Wave DIsplaynewx = $NewSizeWaveX

            NewSizeF = "p" + num2str(n) + "y405_NewF"
            Wave DisplayNewF = $NewSizeF

            //Quick baseline subtraction
            Wavestats DisplayY
            DisplayY -=V_min
            DisplayY /=PowerDensity[2]
            DisplayY *=100000

            DisplayY /= SizeWave[n-1]
            DisplayY /= Exposure405[n-1]

            Wavestats DisplayNew

            DisplayNew -=V_min
            DisplayNew /=PowerDensity[2]
            DisplayNew *=100000

            DisplayNew /= YSize[n-1]
            DisplayNew /= Exposure405[n-1]

            Wavestats DisplayNewX

            DisplayNewX -=V_min
            DisplayNewX /=PowerDensity[2]
            DisplayNewX *=100000

            DisplayNewX /= XSize[n-1]
            DisplayNewX /= Exposure405[n-1]

            Wavestats displaynewF

            DisplayNewF -=V_min
            DisplayNewF /=PowerDensity[2]
            DisplayNewF *=100000

            DisplayNewF /= FSize[n-1]
            DisplayNewF /= Exposure405[n-1]

            AppendToGraph DisplayY vs DisplayX
    EndFor

    Display
    n = 1
    For (n = 1; n <=DimSize(BluePts,0); n += 1)                          // Still need to try to recognize if
the wave exists.

            NameStrY = "p" + num2str(n) + "y350"                // Create a string out of p[n]yCalib, n being
the particle number
            NewSizeWave = "p" + num2str(n) + "y350_New"
            Wave DisplayNew = $NewSizeWave
            NewSizeWaveX = "p" + num2str(n) + "y350_NewX"
            NameStrX = "p" + num2str(n) + "xRef"
            Wave DisplayY = $NameStrY
            Wave DisplayX = $NameStrX
            Wave DIsplaynewx = $NewSizeWaveX

            NewSizeF = "p" + num2str(n) + "y350_NewF"
            Wave DisplayNewF = $NewSizeF

            //Quick baseline subtraction
            Wavestats DisplayY
            DisplayY -=V_min
            DisplayY /=PowerDensity[1]
            DisplayY *=100000

            DisplayY /= SizeWave[n-1]
            DisplayY /= Exposure350[n-1]

            Wavestats DisplayNew
            DisplayNew -=V_min
            DisplayNew /=PowerDensity[1]
            DisplayNew *=100000

            DisplayNew /= YSize[n-1]
            DisplayNew /= Exposure350[n-1]

            Wavestats DisplayNewX
            DisplayNewX -=V_min
            DisplayNewX /=PowerDensity[1]
            DisplayNewX *=100000

            DisplayNewX /= XSize[n-1]
            DisplayNewX /= Exposure350[n-1]

            Wavestats DisplayNewF
            DisplayNewF -=V_min
            DisplayNewF /=PowerDensity[1]
            DisplayNewF *=100000

            DisplayNewF /= FSize[n-1]
            DisplayNewF /= Exposure350[n-1]

            AppendToGraph DisplayY vs DisplayX
    EndFor

    Display

    n = 1
    For(n = 1; n <= DimSize(BluePts,0); n += 1)                          // Still need to try to recognize if
the wave exists.
```

```
                        NameStrY = "p" + num2str(n) + "y280"                    // Create a string out of p[n]yCalib, n being
the particle number
                        NewSizeWave = "p" + num2str(n) + "y280_New"
                        Wave DisplayNew = $NewSizeWave
                        NewSizeWaveX = "p" + num2str(n) + "y280_NewX"
                        NameStrX = "p" + num2str(n) + "xRef"
                        Wave DisplayY = $NameStrY
                        Wave DisplayX = $NameStrX
                        Wave DIsplaynewx = $NewSizeWaveX

                        NewSizeF = "p" + num2str(n) + "y280_NewF"
                        Wave DisplayNewF = $NewSizeF

                        //Quick baseline subtraction
                        Wavestats DisplayY
                        DisplayY -=V_min
                        DisplayY /=PowerDensity[0]
                        DisplayY *=100000

                        DisplayY /= SizeWave[n-1]
                        DisplayY /= Exposure280[n-1]

                        Wavestats DisplayNew
                        DisplayNew -=V_min
                        DisplayNew /=PowerDensity[0]
                        DisplayNew *=100000

                        DisplayNew /= YSize[n-1]
                        DisplayNew /= Exposure280[n-1]

                        Wavestats DisplayNewX
                        DisplayNewX -=V_min
                        DisplayNewX /=PowerDensity[0]
                        DisplayNewX *=100000

                        DisplayNewX /= XSize[n-1]
                        DisplayNewX /= Exposure280[n-1]

                        Wavestats DisplayNewF
                        DisplayNewF -=V_min
                        DisplayNewF /=PowerDensity[0]
                        DisplayNewF *=100000

                        DisplayNewF /= FSize[n-1]
                        DisplayNewF /= Exposure280[n-1]

                        AppendToGraph DisplayY vs DisplayX
        EndFor


        For (n=1; n<=DimSize(AngMom, 0); n+=1)


                PeakSW = "p" + num2str(n) + "y450"
                PeakXRef = "p" + num2str(n) + "xRef"
                PeakY = "p" + num2str(n) + "y450_New"
                PeakX = "p" + num2str(n) + "y450_NewX"
                PeakF = "p" + num2str(n) + "y450_NewF"

                If (Waveexists($PeakSw)==0)
                        Break
```

```
		EndIf

		Wavestats/Q $PeakSW
		PeaksSW[n-1] = V_Max

		Wavestats/Q $PeakY
		PeaksY[n-1] = V_Max

		Wavestats/Q $PeakX
		PeaksX[n-1] = V_Max

		Wavestats/Q $PeakF
		PeaksF[n-1] = V_max

		Wave Degrees

		Make/O FactorSplitY, FactorSplitX

		FactorSplitY = NaN
		FactorSplitX = NaN

		If (Degrees[Up]<90)
			FactorSplitY = Degrees/90
			FactorSplitX  = 1-FactorsplitY
		ElseIf (Degrees[Up]>90)
			FactorSplitY = 90/Degrees
			FactorSplitX  = 1-FactorsplitY
		EndIf
		up += 1
EndFor

For (NaNdel = 1 ; NaNdel <= numpnts(FactorSplitX) ; NaNdel+= 1)
		If (numtype(FactorSplitX[NaNdel-1]) == 2)
			Deletepoints NaNdel-1,1,FactorSplitX
				NaNdel-=1
		Endif
Endfor

For (NaNdel = 1 ; NaNdel <= numpnts(FactorSplitY) ; NaNdel+= 1)
		If (numtype(FactorSplitY[NaNdel-1]) == 2)
			Deletepoints NaNdel-1,1,FactorSplitY
				NaNdel-=1
		Endif
Endfor

For (NaNdel = 1 ; NaNdel <= numpnts(PeaksSW) ; NaNdel+= 1)
		If (numtype(PeaksSW[NaNdel-1]) == 2)
			Deletepoints NaNdel-1,1,PeaksSW
				NaNdel-=1
		Endif
Endfor

For (NaNdel = 1 ; NaNdel <= numpnts(PeaksY) ; NaNdel+= 1)
		If (numtype(PeaksY[NaNdel-1]) == 2)
			Deletepoints NaNdel-1,1,PeaksY
				NaNdel-=1
		Endif
Endfor

For (NaNdel = 1 ; NaNdel <= numpnts(PeaksX) ; NaNdel+= 1)
		If (numtype(PeaksX[NaNdel-1]) == 2)
```

```
                        Deletepoints NaNdel-1,1,PeaksX
                            NaNdel-=1
                Endif
        Endfor

        Edit

        AppendToTable Majmom, Minmom, AspectRatio
End

Function AverageSpectra()
        Make/O/n=301 Avg450, Avg405, Avg350, Avg280

        Wave MajMom
        Variable N
        N=1

        For(n=1;  n<=numpnts(MajMom); N+=1)


                String PeakChange = "p" + num2str(n) + "y450_Final"
                Wave PeakChange1 = $PeakChange
                Avg450 +=PeakChange1

                PeakChange = "p" + num2str(n) + "y405_Final"
                Wave PeakChange1 = $PeakChange
                Avg405 +=PeakChange1

                PeakChange = "p" + num2str(n) + "y350_Final"
                Wave PeakChange1 = $PeakChange
                Avg350 +=PeakChange1

                PeakChange = "p" + num2str(n) + "y280_Final"
                Wave PeakChange1 = $PeakChange
                Avg280 +=PeakChange1

        EndFor

        Avg450 /=Numpnts(Majmom)

End
```

D.2.4 Description of Functions: Code to regrid all emission wavelengths onto 1-nm increments from 400-700 nm.

```
Function RegridF()

Make/O/N=701 Ywv_ref
Make/O/N=701 Xwv_ref
Make/O/N=301 RegriddedX
Make/O/N=301 RegriddedY
Wave Ywave_regrid, BluePts

RegriddedY = NaN

Xwv_ref = 0


String NameStrY, NameStrX, RegridStr, RegridY
```

```
Variable n, h, y

n = 0
h = 200

        For (n = 0; n < DimSize(Xwv_ref, 0); n +=1)
                Xwv_ref[n] = h
                h+=1
        EndFor

h = 400

        For (n = 0; n < DimSize(RegriddedX, 0); n +=1)
                RegriddedX[n] = h
                h+=1
        EndFor


n =1

Ywv_ref = 1

        For (n = 1; n <=DimSize(BluePts,0); n += 1)
                NameStrY = "p" + num2str(n) + "y450_NewF"
                NameStrX = "p" + num2str(n) + "xRef"

                RegridStr = NameStrY + "_regrid"
                regrid_param2($NameStrX, $NameStrY, xwv_ref, ywv_ref)
                duplicate/o ywave_regrid $RegridStr

                For (y=0;y<301;y+=1)
                        RegriddedY[y] = ywave_regrid[y+200]
                EndFor

                RegridY = "p" + Num2Str(n) + "y450_Final"
                Duplicate/o RegriddedY $RegridY

                NameStrY = "p" + num2str(n) + "y405_NewF"
                NameStrX = "p" + num2str(n) + "xRef"

                RegridStr = NameStrY + "_regrid"


                regrid_param2($NameStrX, $NameStrY, xwv_ref, ywv_ref)
                duplicate/o ywave_regrid $RegridStr

                For (y=0;y<301;y+=1)
                        RegriddedY[y] = ywave_regrid[y+200]
                EndFor


                RegridY = "p" + Num2Str(n) + "y405_Final"

                Duplicate/o RegriddedY $RegridY

                NameStrY = "p" + num2str(n) + "y350_NewF"
                NameStrX = "p" + num2str(n) + "xRef"

                RegridStr = NameStrY + "_regrid"
                regrid_param2($NameStrX, $NameStrY, xwv_ref, ywv_ref)
                duplicate/o ywave_regrid $RegridStr

                For (y=00;y<301;y+=1)
                        RegriddedY[y] = ywave_regrid[y+200]
                EndFor

                RegridY = "p" + Num2Str(n) + "y350_Final"

                Duplicate/o RegriddedY $RegridY

                NameStrY = "p" + num2str(n) + "y280_NewF"
                NameStrX = "p" + num2str(n) + "xRef"

                RegridStr = NameStrY + "_regrid"

                regrid_param2($NameStrX, $NameStrY, xwv_ref, ywv_ref)

                duplicate/o ywave_regrid $RegridStr

                For (y=0;y<301;y+=1)
                        RegriddedY[y] = ywave_regrid[y+200]
                EndFor

                RegridY = "p" + Num2Str(n) + "y280_Final"

                Duplicate/o RegriddedY $RegridY

        EndFor
                    // Append the intensity values for each excitation (400-700 nm)
        Edit

        n = 1
        For(n = 1; n <= DimSize(BluePts,0); n += 1)                 // Still need to try to recognize if
the wave exists.
                        NameStrY = "p" + num2str(n) + "y450_Final"          // Create a string out of
p[n]yCalib, n being the particle number
                        AppendToTable $NameStrY
        EndFor

        Edit

        n = 1
        For(n = 1; n <= DimSize(BluePts,0); n += 1)                 // Still need to try to recognize if
the wave exists.
                        NameStrY = "p" + num2str(n) + "y405_Final"          // Create a string out of
p[n]yCalib, n being the particle number
                        AppendToTable $NameStrY
        EndFor

        Edit

        n = 1
        For(n = 1; n <= DimSize(BluePts,0); n += 1)                 // Still need to try to recognize if
the wave exists.
                        NameStrY = "p" + num2str(n) + "y350_Final"          // Create a string out of
p[n]yCalib, n being the particle number
                        AppendToTable $NameStrY
        EndFor

        Edit
    n = 1
    For(n = 1; n <= DimSize(BluePts,0); n += 1)                             // Still need to try to recognize if
the wave exists.
            NameStrY = "p" + num2str(n) + "y280_Final"                      // Create a string out of
p[n]yCalib, n being the particle number
            AppendToTable $NameStrY
    EndFor
End

Function MeasureRegrids()
    Variable n
    Wave bluepts
    Make/O SizeMeasurement_Final, SizeMeasurements
    SizeMeasurement_Final = NaN

    Display

    For (n=1 ;n<=DimSize(BluePts, 0) ;n +=1)

        String SizeWaves = "p" + Num2Str(n) +  "y350_Final"

        Wavestats/Q $SizeWaves
        SizeMeasurement_Final[n-1] = V_Max
        AppendToGraph $SizeWaves
    EndFor

    For (N = 1 ; N <= numpnts(SizeMeasurement_Final) ; N+= 1)
        If (numtype(SizeMeasurement_Final[N-1]) == 2)
            Deletepoints N-1,1,SizeMeasurement_Final
                N-=1
        Endif
    Endfor
End
```

D.2.5 Description of Functions: Plotting for the spectral averages and subsequent deviation of all emission curves for multiple species on a singular graph, as well as the calculations and plotting of box plot parameters for sizing and aspect ratio per species.

```
///////////////// Plot the emission curves as well as deviation for each species //////////////////
Function TestWindows()

    Display
    // De-comment above to create new test graph! //

    Wave textWave0
    Variable fracVar, windowVar, n, j, YAxisUpVar, YAxisDownVar, AStorageVar, AxisVar
    String WaveCheckStr, AvgYStr, AvgXStr, YAxisStr, XAxisStr, StDevY, StDevU, StDevL

    AStorageVar = 0

    fracVar = numpnts(textWave0)
    print fracVar

    windowVar = 100/fracVar
    print windowVar

    //////////////////////////////////////////////////////////// 450
```

```
j = 1

For(n=1;j==1;n+=1)

        AvgYStr = "y" + num2str(n) + "d"
        AvgXStr = "yXd"

        StDevY = "y" + num2str(n) + "d_s"
        String StDevY2 = "y" + num2str(n) + "d_s_COPY"

        StDevU = "y" + num2str(n) + "d+sU"
        StDevL = "y" + num2str(n) + "d+sL"

        Make/O/N=301 $StDevU
        Make/O/N=301 $StDevL

        Wave UpperSD = $StDevU
        Wave LowerSD = $StDevL

        Duplicate/O $StDevY $StDevY2

        String AvgYStr2 = "y" + num2str(n) + "d_COPY"

        Duplicate/O $AvgYStr $AvgYStr2

        Wave AverageWv = $AvgYStr2
        Wave StDevWv = $StDevY2

        UpperSD = AverageWv + StDevWv
        LowerSD = AverageWv - StDevWv

        YAxisStr = "y" + num2str(n) + "d"
        XAxisStr = "y" + num2str(n) + "dx"

        YAxisUpVar = (windowVar*n)/100
        YAxisDownVar = ((windowVar*n)-10)/100

        Wavestats $AvgYStr

        If(V_Max > AStorageVar)
                AStorageVar = V_Max
        EndIf

        AppendToGraph/L=$YAxisStr/B=$XAxisStr $AvgYStr vs $AvgXStr
        ModifyGraph axisEnab($YAxisStr)={YAxisDownVar,YAxisUpVar},axisEnab($XAxisStr)={0.25,0.35}
        ModifyGraph freePos($YAxisStr)={0,$XAxisStr},freePos($XAxisStr)={0,$YAxisStr}

        WaveCheckStr = "y" + num2str(n+1) + "d"
        j = WaveExists($WaveCheckStr)

        ModifyGraph nticks($XAxisStr)=0
        ModifyGraph lsize($AvgYStr )=2,rgb($AvgYStr )=(65535,0,52428)

        ModifyGraph manTick($AvgYStr)={0,14,0,0},manMinor($AvgYStr)={0,0}
Endfor

j=1

AxisVar = AStorageVar/3

For(n=1;j==1; n+=1)

        YAxisStr = "y" + num2str(n) + "d"

        SetAxis $YAxisStr 0,AStorageVar
        ModifyGraph manTick($XAxisStr)={0,AxisVar,0,0},manMinor($XAxisStr)={0,0}

        WaveCheckStr = "y" + num2str(n+1) + "d"
        j = WaveExists($WaveCheckStr)
        ModifyGraph freePos($YAxisStr)={400,$XAxisStr}
    EndFor

    //////////////////////////////////////////////////////////// 405

    AStorageVar = 0
    j = 1

    For(n=1;j==1;n+=1)

        AvgYStr = "y" + num2str(n) + "c"
        AvgXStr = "yXc"

        YAxisStr = "y" + num2str(n) + "c"
        XAxisStr = "y" + num2str(n) + "acx"

        YAxisUpVar = (windowVar*n)/100
        YAxisDownVar = ((windowVar*n)-10)/100

        Wavestats $AvgYStr

        If(V_Max > AStorageVar)
            AStorageVar = V_Max
        EndIf

        AppendToGraph/L=$YAxisStr/B=$XAxisStr $AvgYStr vs $AvgXStr
        ModifyGraph
axisEnab($YAxisStr)={YAxisDownVar,YAxisUpVar},axisEnab($XAxisStr)={0.42,0.57};DelayUpdate
        ModifyGraph freePos($YAxisStr)={0,$XAxisStr},freePos($XAxisStr)={0,$YAxisStr}

        WaveCheckStr = "y" + num2str(n+1) + "c"
        j = WaveExists($WaveCheckStr)

        ModifyGraph nticks($XAxisStr)=0
        ModifyGraph lsize($AvgYStr)=2,rgb($AvgYStr)=(26411,1,52428)

        ModifyGraph manTick($AvgYStr)={0,0.4,0,1},manMinor($AvgYStr)={0,0}
    Endfor

    j=1

    AxisVar = AStorageVar/3

    For(n=1;j==1; n+=1)

        YAxisStr = "y" + num2str(n) + "c"

        SetAxis $YAxisStr 0,AStorageVar

        ModifyGraph manTick($XAxisStr)={0,AxisVar,0,0},manMinor($XAxisStr)={0,0}

        WaveCheckStr = "y" + num2str(n+1) + "c"
        j = WaveExists($WaveCheckStr)
        ModifyGraph freePos($YAxisStr)={400,$XAxisStr}
    EndFor

    //////////////////////////////////////////////////////////// 350
    AStorageVar = 0
    j = 1
    For(n=1;j==1;n+=1)

        AvgYStr = "y" + num2str(n) + "a"
        AvgXStr = "yXa"

        YAxisStr = "y" + num2str(n) + "a"
        XAxisStr = "y" + num2str(n) + "aax"

        YAxisUpVar = (windowVar*n)/100
        YAxisDownVar = ((windowVar*n)-10)/100

        Wavestats $AvgYStr

        If(V_Max > AStorageVar)
            AStorageVar = V_Max
        EndIf

        AppendToGraph/L=$YAxisStr/B=$XAxisStr $AvgYStr vs $AvgXStr
        ModifyGraph
axisEnab($YAxisStr)={YAxisDownVar,YAxisUpVar},axisEnab($XAxisStr)={0.82,.97};DelayUpdate
        ModifyGraph freePos($YAxisStr)={0,$XAxisStr},freePos($XAxisStr)={0,$YAxisStr}

        WaveCheckStr = "y" + num2str(n+1) + "a"
        j = WaveExists($WaveCheckStr)

        ModifyGraph nticks($XAxisStr)=0
        ModifyGraph lsize($AvgYStr)=2,rgb($AvgYStr)=(1,52428,26586)
        SetAxis $XAxisStr 400,700

        ModifyGraph manTick($AvgYStr)={0,7,0,0},manMinor($AvgYStr)={0,0}
    Endfor

    j=1

    AxisVar = AStorageVar/3

    For(n=1;j==1; n+=1)

        YAxisStr = "y" + num2str(n) + "a"

        SetAxis $YAxisStr 0,AStorageVar

        ModifyGraph manTick($XAxisStr)={0,AxisVar,0,0},manMinor($XAxisStr)={0,0}

        WaveCheckStr = "y" + num2str(n+1) + "a"
        j = WaveExists($WaveCheckStr)
        ModifyGraph freePos($YAxisStr)={400,$XAxisStr}
    EndFor

    //////////////////////////////////////////////////////////////// 280
    AStorageVar = 0
    j = 1

    For(n=1;j==1;n+=1)

        AvgYStr = "y" + num2str(n) + "b"
        AvgXStr = "yXb"

        YAxisStr = "y" + num2str(n) + "b"
        XAxisStr = "y" + num2str(n) + "abx"

        YAxisUpVar = (windowVar*n)/100
        YAxisDownVar = ((windowVar*n)-10)/100
```

```
                Wavestats $AvgYStr

                If(V_Max > AStorageVar)
                        AStorageVar = V_Max
                EndIf

                AppendToGraph/L=$YAxisStr/B=$XAxisStr $AvgYStr vs $AvgXStr
                ModifyGraph
axisEnab($YAxisStr)={YAxisDownVar,YAxisUpVar},axisEnab($XAxisStr)={.62,0.77};DelayUpdate
                ModifyGraph freePos($YAxisStr)={0,$XAxisStr},freePos($XAxisStr)={0,$YAxisStr}

                WaveCheckStr = "y" + num2str(n+1) + "b"
                j = WaveExists($WaveCheckStr)

                ModifyGraph nticks($XAxisStr)=0
                ModifyGraph lsize($AvgYStr)=2,rgb($AvgYStr)=(0,0,65535)

                ModifyGraph manTick($AvgYStr)={0,6,0,0},manMinor($AvgYStr)={0,0}
        Endfor

        j=1

        AxisVar = AStorageVar/3

        For(n=1;j==1; n+=1)

                YAxisStr = "y" + num2str(n) + "b"

                SetAxis $YAxisStr 0,AStorageVar

                ModifyGraph manTick($XAxisStr)={0,AxisVar,0,0},manMinor($XAxisStr)={0,0}

                WaveCheckStr = "y" + num2str(n+1) + "b"
                j = WaveExists($WaveCheckStr)
                ModifyGraph freePos($YAxisStr)={400,$XAxisStr}
        EndFor

                ModifyGraph fSize=16

End

////////////////// Calculate and plot box-plots for size and aspect ratio (per species) //////////////////

Function NewBoxPlots()

        Variable j, n, V_Q25, V_Q75, V_Median, V_IQR, Step, lowerInnerFence, LowerOuterFence, upperinnerfence,
upperouterfence
        String xStr, WaveExistsStr, typeStr, NewMedianStr
        Wave P_F1_50, P_F1_25, P_F1_75, P_F1_10, P_F1_90


        n = 1

        WaveExistsStr = "s" + Num2Str(n) + "maj"

        Make/O/T ListWv

        For(n=1;WaveExists($WaveExistsStr)== 1; n+=1)

                Sort/R $WaveExistsStr, $WaveExistsStr

                WaveExistsStr = "s" + Num2Str(n) + "maj"

                ListWv[n-1] = WaveExistsStr
        EndFor

        For (n = 0 ; n <= numpnts(ListWv) ; n+= 1)
                If (strlen(ListWv[n]) == 0)
                        Deletepoints n,1,ListWv
                        n-=1
                Endif
        Endfor

        Print ListWv[0]

        String ListStr = ListWv[0]

        For(n = 1 ; n < numpnts(ListWv) ; n+= 1)

                ListStr += ";" + ListWv[n]

        EndFor

        Print ListStr

        Make/O/N=(n) MediansWv
        Make/O/N=(n) Q_25
        Make/O/N=(n) Q_75
        Make/O/N=(n) Q_10
        Make/O/N=(n) Q_90
        Make/O/N=(n) StepWv

        Q_25 = NaN


        n = 1
        WaveExistsStr = "s" + Num2Str(n) + "maj"

//      fWavePercentile(StringFromList(0,ListStr, ";"), "10;25;50;75;90", "P_F1",0,1,1.5)
        Make/O/N=(numpnts(ListWv)) IQR

        For(n=1;n<=Numpnts(ListWv); n+=1)

                String OutlierUStr
                String OutlierDStr

                Make/O $OutlierUStr
                Make/O $OutlierDStr

                WaveExistsStr = "s" + Num2Str(n) + "maj"
        //      StatsQuantiles/BOX $WaveExistsStr

                fWavePercentile(StringFromList(n-1,ListStr, ";"), "10;25;50;75;90", "P_F1",0,0,0)

                MediansWv[n-1] = P_F1_50
                Q_25[n-1] = P_F1_25
                Q_75[n-1] = P_F1_75

                IQR[n-1] = Q_75[n-1]-Q_25[n-1]
                StepWv[n-1] = IQR*1.5
                Q_10[n-1] = P_F1_10
                Q_90[n-1] = P_F1_90
        EndFor

        Variable p = 0

        Make/O OutlierWv
        OutlierWv = 0
```

```
Make/O/T OutlierStrWv
OutlierStrWv = "Fix Needed"

For(n=1;n<=Numpnts(ListWv); n+=1)
    WaveExistsStr = "s" + Num2Str(n) + "maj"

    For(j=1; j<=Numpnts($WaveExistsStr);j+=1)

        Wave TestWv = $WaveExistsStr

        If(TestWv[j-1]>(Q_75[n-1]+StepWv[n-1]) )
    //  If(TestWv[j-1]>5 )
            print 1
            String OutlierStr = WaveExistsStr + "_OL"

            OutlierWv[p] +=TestWv[j-1]
            wave OutlierWV = $OutlierStr

            OutlierStrWv[p] = WaveExistsStr
            p+=1
        ElseIf(TestWv[j-1]<(Q_25[n-1]-StepWv[n-1]))
            OutlierStr = WaveExistsStr + "_OL"

            OutlierWv[p] +=TestWv[j-1]
            wave OutlierWV = $OutlierStr

            OutlierStrWv[p] = WaveExistsStr
            p+=1
        Else

        EndIf

    EndFor

EndFor

For (n = 0 ; n <= numpnts(OutlierWV) ; n+= 1)
    If (OutlierWV[n] == 0)
        Deletepoints n,1,OutlierWV
        n-=1
    Endif
Endfor

For (n = 0 ; n <= numpnts(OutlierStrWv) ; n+= 1)
    If (StringMatch("Fix Needed", OutlierStrWv[n]))
        Deletepoints n,1,OutlierStrWv
        n-=1
    Endif
Endfor

For (n = 0 ; n <= numpnts(MediansWv) ; n+= 1)
    If (MediansWv[n] == 0)
        Deletepoints n,1,MediansWv
        Deletepoints n,1,ListWv
        DeletePoints n,1,StepWv
        Deletepoints n,1,Q_90
        Deletepoints n,1,Q_75
        Deletepoints n,1,Q_25
        Deletepoints n,1,Q_10
        n-=1
    Endif
Endfor

Make/O/N=(numpnts(Q_90)) Q_25P
Make/O/N=(numpnts(Q_90)) Q_75P
```

```
Q_75P = (Q_75-MediansWv)
Q_25P = (MediansWv-Q_25)

Make/O/N=(numpnts(Q_90)) Q_10P
Make/O/N=(numpnts(Q_90)) Q_90P

Q_90P = (Q_90-Q_75)
Q_10P = (Q_25-Q_10)

String SingleStr, SingleStr2
Variable y

For(n=0;n<=Numpnts(MediansWv);n+=1)

        SingleStr = "s" + num2str(n+1) + "_10"
        Wave RefWv = $SingleStr
        RefWv = Q_10[n]

        SingleStr = "s" + num2str(n+1) + "_25"
        Wave RefWv = $SingleStr
        RefWv = Q_25[n]

        SingleStr = "s" + num2str(n+1) + "_50"
        Wave RefWv = $SingleStr
        RefWv = MediansWv[n]

        SingleStr = "s" + num2str(n+1) + "_75"
        Wave RefWv = $SingleStr
        RefWv = Q_75[n]

        SingleStr = "s" + num2str(n+1) + "_90"
        Wave RefWv = $SingleStr
        RefWv = Q_90[n]

        SingleStr2 = "s" + num2str(n+1) + "_75P"
        Make/O/N=1 $SingleStr2
        Wave RefWv = $SingleStr2
        RefWv = Q_75P[n]

        SingleStr2 = "s" + num2str(n+1) + "_25P"
        Make/O/N=1 $SingleStr2
        Wave RefWv = $SingleStr2
        RefWv = Q_25P[n]

        SingleStr2 = "s" + num2str(n+1) + "_10P"
        Make/O/N=1 $SingleStr2
        Wave RefWv = $SingleStr2
        RefWv = Q_10P[n]

        SingleStr2 = "s" + num2str(n+1) + "_90P"
        Make/O/N=1 $SingleStr2
        Wave RefWv = $SingleStr2
        RefWv = Q_90P[n]

        SingleStr = "s" + num2str(n+1) + "maj"
        String OutliersStr = "s" + num2str(n+1) + "_OL"

        Make/O/N=(numpnts(OutlierWv)) RefOlWv
        wave RefOlWv = $OutliersStr

        Redimension/N=(numpnts(OutlierWv)) RefOlWv

                For(y=1;y<=numpnts(OutlierWv);y+=1)
                        If(stringmatch(SingleStr, OutlierStrWv[y-1])==1)
```

```
                    RefOlWv[y-1]=OutlierWv[y-1]
                EndIf
            EndFor

    EndFor

    //////////////////////////////////////////////////
    //******** Inter Code Split ************//
    //          Append To Graph Now         //
    //******** Inter Code Split ************//
    //////////////////////////////////////////////////

    // NEED TO MAKE THIS A LOOP AND REPLACE SizeY AXIS LABEL
    Make/O/N=1 PositioningWv
    PositioningWv = 1

    Make/O/N=1 PositioningWv2
    PositioningWv2 = 2

    Variable fracVar, windowVar, YAxisUpVar, YAxisDownVar

    fracVar = numpnts(ListWv)

    windowVar = 100/fracVar
    Wave textWave0

For(n=0; n<=numpnts(ListWv); n+=1)

    YAxisUpVar = (windowVar*(n+1))/100
    YAxisDownVar = ((windowVar*(n+1))-10)/100

    String S10PStr = "s" + num2str(n+1) + "_10p"
    String S25Str = "s" + num2str(n+1) + "_25P"
    String S50Str = "s" + num2str(n+1) + "_50"
    String S75Str = "s" + num2str(n+1) + "_75P"
    String S90PStr = "s" + num2str(n+1) + "_90p"
    String SOLStr = "s" + num2str(n+1) + "_OL"
    String S90Str = "s" + num2str(n+1) + "_90"
    String S10Str = "s" + num2str(n+1) + "_10"

    String XAxisStr = "s" + num2str(n+1) + "x"
    String YAxisStr = "s" + num2str(n+1) + "y"


    AppendToGraph/L=$YAxisStr/B=$XAxisStr $S50Str vs PositioningWv                  // Position for median

    SetAxis $XAxisStr 0,3
    SetAxis $YAxisStr 20,120
    ModifyGraph mode=2;DelayUpdate
    ErrorBars $S50Str X,const=0.3                                                      // Median
line


    String S50_Copy = "s" + num2str(n+1) + "_50#1"

    ErrorBars/L=0 $S50_Copy BOX,const=0.3,wave=($S75Str,$S25Str)

    //~~~

    Wave RefWv = $SOLStr

    If(RefWv ==0)
        AppendToGraph/L=$YAxisStr/B=$XAxisStr RefWv vs PositioningWv
    EndIf
```

```
    ModifyGraph axisEnab($YAxisStr)={YAxisDownVar,YAxisUpVar},axisEnab($XAxisStr)={0.07,0.13}
    ModifyGraph freePos($YAxisStr)={0,$XAxisStr},freePos($XAxisStr)={0,$YAxisStr}
    ModifyGraph freePos($YAxisStr)={20,$YAxisStr}

    ModifyGraph mode($SOLStr)=3,marker($SOLStr)=8,msize($SOLStr)=2

    AppendToGraph/L=$YAxisStr/B=$XAxisStr $S50Str vs PositioningWv                    // Position for box
    ErrorBars/L=0 $S50_Copy BOX,const=0.3,wave=($s75str,$s25str) // Make that box

    AppendToGraph/L=$YAxisStr/B=$XAxisStr $s90str vs PositioningWv                    // 90th percentile line
    ErrorBars/T=0 $s90Str XY,const=0.3,wave=(,$s90Pstr)          // 90th whisker

    AppendToGraph/L=$YAxisStr/B=$XAxisStr $s10str vs PositioningWv          //10th line
    ErrorBars/T=0 $s10Str XY,const=0.3,wave=($s10Pstr,)          // 10th whisker

    ModifyGraph freePos($YAxisStr)={0,$XAxisStr},freePos($XAxisStr)={20,$YAxisStr}
    ModifyGraph nticks($XAxisStr)=2,userticks($XAxisStr)={MediansWv,textWave0}

    ErrorBars/T=0 $S50Str X,const=0.3

    ModifyGraph rgb($S50Str)=(0,0,0),rgb($S50_Copy)=(0,0,0),rgb(s1_50#2)=(0,0,0);
    ModifyGraph rgb($S90Str)=(0,0,0),rgb($S10Str)=(0,0,0)

    String RemoveLast = "s" + num2str(n) + "_50#2"
    RemoveFromGraph $RemoveLast

    ModifyGraph manTick($YAxisStr)={20,50,0,0},manMinor($YAxisStr)={0,0}
    ModifyGraph manTick($YAxisStr)={20,50,0,0},manMinor($YAxisStr)={5,0}
EndFor
```

D.2.6 Description of Functions: Time-resolution code for plotting and analysis of Cyprus and AQABA data sets after the Random Forest classification.

```
Function plotTRData()
    Wave ParticleTime, Type, AmbDust, RiboB, ChloroB, TrypB, BacB, NADB, TypeNum
    Variable B,E, n, Q, m, nandel

    Make/O/n=300000 AmbDust, RiboB, ChloroB, TrypB, BacB, NADB

    AmbDust = 0
    RiboB = 0
    ChloroB = 0
    TrypB = 0
    BacB = 0
    NADB = 0
    Q=0
    B=ParticleTime[Q]
    E=B+300 //////////////////////// Change this number (in seconds) to change time resolution 300 = 5 mins //////////////////////

    n=0
    m=0

    Make/O/D/N=300000 WaveTime
    WaveTime = 0
    WaveTime[m] = ParticleTime[m]

    For(n=0; n<=numpnts(ParticleTime); n+=1)
        E=B+300

        if(ParticleTime[n]<=E)
            if(TypeNum[n]==1)
                AmbDust[m]+=1
            Elseif(TypeNum[n]==2)
                NADB[m]+=1
            Elseif(TypeNum[n]==3)
                RiboB[m]+=1
            ElseIf(TypeNum[n]==4)
                ChloroB[m]+=1
            ElseIf(TypeNum[n]==5)
                TrypB[m]+=1
            ElseIf(TypeNum[n]==6)
                BacB[m]+=1
            Endif
        Else
            B+=300
            m+=1
            WaveTime[m] = B
            print "Five Min"
        EndIf
        print n
    EndFor
End
```