Investigating Aerosol Composition with Low Cost Optical Particle Counters

by

Will Sharpe

B.S., Aerospace Engineering
University of Maryland, 2019

Submitted to the Department of Civil and Environmental Engineering
in partial fulfillment of the requirements for the degree of

Master of Engineering in Civil and Environmental Engineering

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2023

Signature of Author . . . .
Will Sharpe
Department of Civil and Environmental Engineering
May 12, 2023

Certified by: . . . .
Jesse Kroll
Department of Civil and Environmental Engineering, Thesis Supervisor

Accepted by: . . . .
Colette L. Heald
Professor of Civil and Environmental Engineering
Chair, Graduate Program Committee

Investigating Aerosol Composition with Low Cost Optical Particle Counters

by

Will Sharpe

B.S., Aerospace Engineering
University of Maryland, 2019

Submitted to the Department of Civil and Environmental Engineering
in partial fulfillment of the requirements for the degree of

Master of Engineering in Civil and Environmental Engineering

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2023

Authored by: Will Sharpe
Department of Civil and Environmental Engineering
May 12, 2023

Certified by: Jesse Kroll
Department of Civil and Environmental Engineering
Thesis Supervisor

Accepted by: Colette L. Heald
Professor of Civil and Environmental Engineering
Chair, Graduate Program Committee

Investigating Aerosol Composition with Low Cost Optical Particle Counters

by

Will Sharpe

Submitted to the Department of Civil and Environmental Engineering
on May 12, 2023, in partial fulfillment of the
requirements for the degree of
Master of Engineering in Civil and Environmental Engineering

ABSTRACT

Particulate matter (PM) is a serious threat to human health and contributes to millions of premature deaths a year globally. Access to source attribution and compositional data of PM can have many benefits from easier regulation to enabling a better understanding of the negative health effects associated with PM. Acquiring compositional data for ambient PM generally has a high associated cost and is done using complex instrumentation, manual postprocessing, and labor intensive lab work. These approaches produce very high quality data, but have low spatiotemporal resolution and a high cost. This work explores a novel method to generate basic compositional data for ambient PM with low-cost, easily deployable apparatuses in concert with a simple fully connected neural net. Simulated effects of thermal denuders as well as dryers/humidifiers are used to perturb aerosols before they enter simulated low-cost optical particle counters (OPCs). This provides information on the volatility and hygroscopicity of the aerosols. These OPC outputs are processed programmatically and fed into a neural net to classify what category an incoming aerosol belongs to. This method is run for both compound-derived categories which mimic real PM sources (Sea Salt, Biomass Burning, Dust, and Urban Smog), and property-derived aerosols which present more idealized conditions. The results of this method are near-perfect classification for single mode aerosol distributions and over 90% correct classification for two mode aerosol distributions. The results on the property-derived aerosols have shown robustness to changing aerosol properties, as well as to changing apparatus and ambient conditions. This work provides proof of concept for future real world experiments to verify this method and presents an experimental setup for this purpose. Having access to compositional data for ambient PM should allow access to PM sources at a very high spatiotemporal resolution for a relatively low price. This basic source attribution could provide the data needed for better informed regulation as well as future scientific work.

Thesis Supervisor: Jesse Kroll
Title: Professor of Civil and Environmental Engineering and Chemical Engineering

Acknowledgments

I would like to express my sincere gratitude to my thesis advisor, Dr. Jesse Kroll, for his guidance, support, and patience throughout my research project. His insights and expertise were invaluable in shaping my research questions and methodology.

I would also like to thank my fellow lab members for answering the many questions I had about both MIT in general and the specific workings of our lab. Without their help I would have been lost in all my experimental work.

Finally I would like to thank David Hagan and QuantAQ for the use of their instruments and for providing valuable insight on the specifics of these instruments and their best practices.

THIS PAGE INTENTIONALLY LEFT BLANK

Contents

List of Figures

List of Tables

THIS PAGE INTENTIONALLY LEFT BLANK

Chapter 1

Introduction

1.1 Particulate Matter

Small particulate matter in the air, especially particles smaller than 2.5 μm ($PM_{2.5}$) can have severe impacts on human health. These particles are small enough to travel through the respiratory tract reaching the lungs, and entering the circulatory system [1]. This can lead to an array of negative health effects from pulmonary and cardiovascular disease to adverse birth outcomes [2]. Models estimated the premature mortality from $PM_{2.5}$ at 8.9 million in 2015 [3], because of this massive impact $PM_{2.5}$ is widely studied and monitored. There are multiple layers of fidelity one can hope to achieve when monitoring PM. The simplest methods are counting the total number of particles, or the total mass, which are able to be achieved with relatively simple instruments. A harder problem is determining PM sources and composition.

There are a very wide range of PM sources globally, and local PM concentrations depend mainly on local factors [4]. PM can come from both natural sources such as wildfires or dust and anthropogenic sources such as fuel combustion and agriculture. "Secondary" particulate matter can also be formed away from a PM source via reaction with other pollutants in the air such as sulfur dioxide or ammonia [5]. Since precise global source attribution would be an impossible task, there has been substantial work into identifying general categories of ambient PM sources. One meta-analysis of over 200 published works on PM source attribution found six major categories:

Sea/Road Salt, Crustal/Mineral Dust, Secondary Inorganic Aerosols, Traffic (SOAs), Point Sources and Biomass Burning [6]. A WHO global meta analysis grouped ambient aerosol sources by Traffic, Industry, Domestic fuel burning, Natural Sources including Sea Salt and Dust, and an unspecified category [7]. A more recent global study built upon that WHO work and identified Secondary Inorganic Species, Sea Salt, Dust, Traffic, Industry, Biomass Burning, Coal/Oil combustion, and Other as its categories [8].

These different particle sources can potentially lead to differing negative health outcomes. A WHO study explored the differing negative health effects of various aerosol sources, and suggested a difference in health outcomes between aerosols generated from biomass burning, coal combustion, and dust. This same study reviewed literature indicating even short term high level exposures to combustion-derived particles can lead to immediate adverse health effects [9]. Another study pointed to traffic sources as the main driver for adverse health effects associated with ambient $PM_{2.5}$ [10]. Many similarly scoped studies have been performed on the differing health effects from different sources, but it is important to note that this is a field of active research and how confident one can be in these studies with the current data available is currently contested [11]. This lack of available data is hampering the ability to get a conclusive mapping of health effects to PM sources. Having this conclusive mapping between PM sources and health effects would have a tremendous benefit and would allow for more effective policy decisions, higher fidelity risk area mapping, and improved ability to abate negative health outcomes. Helping to ameliorate this data limitation could enable further research and have substantial positive impact.

A second major benefit of having an understanding of particle sources is easier regulation. If a policy maker only knows the total amount of PM in their area, regulating it could be a very difficult task and one might have to rely on more general principles. If the sources of PM are known the task of regulation becomes much more straightforward, particularly if this information has high spatial resolution. There are resources such as the EPA's Air Pollution Cost Control Manual, and Control Strategy Tool which detail effective ways to control various pollutants including PM for a particular budget. But many of the methods detailed by these tools are for point sources of pollution and depend on the characteristics of the particles [12]. With higher spatial resolution a policy maker would know more precisely where the biggest point sources of pollution are, and with basic compositional data they would be able to better apply the existing control strategies.

1.2 Traditional Particulate Matter Composition Measurement

In the US the major networks collecting PM composition data and performing source attribution are the EPA's Chemical Speciation Network, and the IMPROVE network [13]. These networks consist of hundreds of monitoring sites spread across the US which measure the major chemical components of $PM_{2.5}$ and eventually produce a large, publicly available dataset. These networks use daily and hourly averages and are able to produce near complete mass balances. This is achieved using collected data from filters in conjunction with chemical lab analysis done at a later date. The end product is high quality compositional data, but there are a few drawbacks. Since this data is only collected at specified sites it has low spatial resolution. One can determine a very well defined picture of PM composition over time at the site locations, but that does not necessarily translate to regions not in the immediate vicinity of a site. The second drawback is the time needed to generate these results. Lab analysis, as well as using daily or hourly averages means that data is not immediately available, and does not have high temporal resolution. A final drawback is the substantial cost associated with both the instruments used and personnel required to run each of these sites.

Additional methods for obtaining compositional data from aerosols do exist outside of site based monitoring. Instruments such as the Aerosol Chemical Speciation Monitor from Aerodyne Research allow for automated compositional data with a 1 minute time resolution, and do not require the same level of laboratory analysis [14]. The trade off versus site based monitoring is that these instruments do not measure for as many compounds as the EPA sites. Instruments such as these provide solutions to some of the temporal problems, but they still suffer from some of the same problems as site based monitoring. These instruments are still relatively expensive and not trivially deployable. These factors contribute to an overarching difficulty in generating a high resolution map of PM composition.

1.3 Optical Particle Counters

In recent years low-cost optical particle counters (OPCs) have been increasingly researched due to their ability to provide relatively accurate air quality data, including $PM_{2.5}$ counts, at a very low price point. OPCs measure incoming particles using a laser of known wavelength and light scattering, where the amount of scattered light is associated with a particle's size using Mie theory. Mie theory, developed by Gustav Mie, provides an exact solution for Maxwell's Equations for a spherical particle [15]. For a given spherical particle with known refractive index, incident light of known wavelength, and known scattering of that light (Mie scattering), Mie theory allows computation of that spherical particle's radius. These sensors use Mie theory to individually detect each particle and output a binned size count for the incoming aerosol distribution [16]. Since low-cost OPCs do not have the ability to measure a particle's refractive index (a required input to use Mie Theory), OPCs are factory calibrated to a certain index. This introduces a limitation that OPCs would be expected to produce less accurate results for particles with refractive indices very far away from the calibration index. This work will be focusing on use of the Alphasense N3 OPC. This OPC has 24 size bins and can detect particles from 0.35 μm to 40 μm. This particular OPC is calibrated to particles with a refractive index of $1.5 + 0j$ [17].

Low-cost OPCs, generally priced around $200-$2000, have the potential to solve some of the problems of traditional PM compositonal determination methods. OPCs are both low-cost and easily deployable. This affords the ability to disperse these sensors more thoroughly than you could a higher cost instrument or station. OPCs also sample relatively rapidly (on the order of seconds) and are thus able to achieve high temporal resolution. While these OPCs solve some spatiotemporal problems, their main drawbacks are their relative inaccuracy compared to site based monitoring, and that they are only able to count the total number of particles in each size bin.

1.4 Determining PM Composition with Low-Cost OPCs

This work outlines a method to get information about aerosol composition and sources using low-cost OPCs. This approach allows access to information about pollution sources in addition to pollution amounts at a very low price point. Instead of taking on the problem of trying to determine exactly what an incoming aerosol distribution consists of, this work aims to estimate accessible properties of aerosols. The properties chosen for this work are hygroscopicity and volatility, since they are fairly easy to measure with low-cost apparatuses.

Low-cost OPCs can be substantially influenced by outside factors such as temperature and relative humidity, which can lead to a major change in output if left unaccounted for [18]. The methods shown in this work will attempt to gain additional compositional information about aerosols using these influences. This will be done by using thermal denuders (TDs), dryers, and humidifiers to modify temperature and relative humidity (RH). Experiments will be simulated in which aerosol distributions pass through varied temperature and RH conditions before being collected by low-cost OPCs. This data will then be processed and used to algorithmicly categorize the incoming distribution via a neural network into either an exemplar category or urban smog, dust, biomass burning, or sea salt aerosol. This would allow access to a level of compositional data without additional human effort at a very low price point and is discussed in depth in the methodology section.

This work will focus on four aerosol categories: Dust, Sea Salt, Biomass Burning, and Urban Smog. These categories are defined to be associated with different aerosols sources. These categories are necessarily broad, and capture different region's primary aerosol sources, a wide range of aerosol compositions, and a wide range of particle sizes. In addition, the approach taken in this work is category agnostic, and could be applied to other category choices one might make. More in depth discussion of the chosen categories is available in the methodology section.

1.5 Existing Research

1.5.1 Low-Cost OPCs

The basis for both the theoretical and software work in this project comes from past work of MIT's Kroll Group. Opcsim is a software package developed by David Hagan to simulate the response of OPCs as well as Nephelometers given a specific set of inputs [16]. This same work presented an overview of the problems associated with low-cost OPCs as well as assumptions in their code. Additionally the use of low-cost apparatuses, including an Alphasense OPC, to explore aerosol sources has been researched experimentally by Kroll group members in the past [19].

Alphasense OPCs have been found to have relatively accurate outputs, particularly for particles above 0.8 μm (83-101% of true value) [20]. This accuracy tends to degrade as particles get smaller. Calibration and RH have been shown to substantially affect Alphasense OPC precision [21]. The importance of calibration for these OPCs is something that must be kept in mind. Incorrect calibration could lead to errors in the output were someone to construct a physical array. Work has been done on the calibration of these low-cost sensors [22], which helps to show and quantify these potential errors.

1.5.2 Aerosol Composition and Categorization

Another important area of this work is aerosol categorization. One paper on categorizing aerosol composition using size distributions and hygroscopicity has been done previously. This paper did not account for volatility but was able to gain basic information about an incoming aerosol's composition via the size growth at differing

RHs [23]. There has also been research on using a thermal denuder in concert with other apparatuses to measure aerosol volatility [24], and research suggesting that different PM sources have distinct volatility fractions [25]. No research was found that used both volatility and hygroscopicity in parallel.

The use of machine learning to classify aerosols is currently sparsely studied. Only one paper was found on this topic, and it was focused on single-particle mass spectrometry measurements [26]. But machine learning for classification problems has been studied very widely. The code for classification makes use of the AutoML mljar-supervised package which contains low code solutions for classification [27] as well as custom fully connected neural networks. These neural network are built using the Keras python package which is widely used in both industry and academia [28].

1.5.3 Paper Organization

This section presents the questions to be solved as well as background and motivation. Chapter 2 will first focus on existing software and underlying equations used to model the physical processes happening to the aerosols. Then it will cover how OPC outputs are processed, how machine learning is used to extract compositional information from these aerosol distributions, and finally a basic labratory setup. Chapter 3 will go through the results of applying these methods first to single mode aerosol distributions then to two mode distributions. Sensitivity of the model is then discussed to help motivate future construction of experimental arrays. Finally the impact of this work and the opportunities for future work are discussed in the Chapter 4.

THIS PAGE INTENTIONALLY LEFT BLANK

Chapter 2

Methodology

2.1 Methodology Overview

This sections focuses on presenting a software architecture that simulates the output of an array of OPCs to aerosols which have been passed through either a dryer, humidifier, TD or a combination of the three. This architecture then processes that data and then passes it to a neural network (NN). A high level overview of this software architecture is shown in the Figure 2-1. Five different experimental conditions are simulated for the data generation with each datapoint containing information from all five experimental conditions. After the outputs from these OPCs are simulated they are processed in one of four ways to help make the data more suitable for a simple neural network. Once the data is processed a neural network can be trained and used to predict the classes of unseen aerosols. Figure 2-1 also show how a future user could use the trained NN with a real world array. The OPCs which this work has been built around output their results to csv files and code has been developed to take in five of these files (one for each experimental condition in the image) and categorize the aerosols into one of the categories discussed below. To help facilitate the creation of this array in the real world a simple laboratory setup is presented at the end of this section.

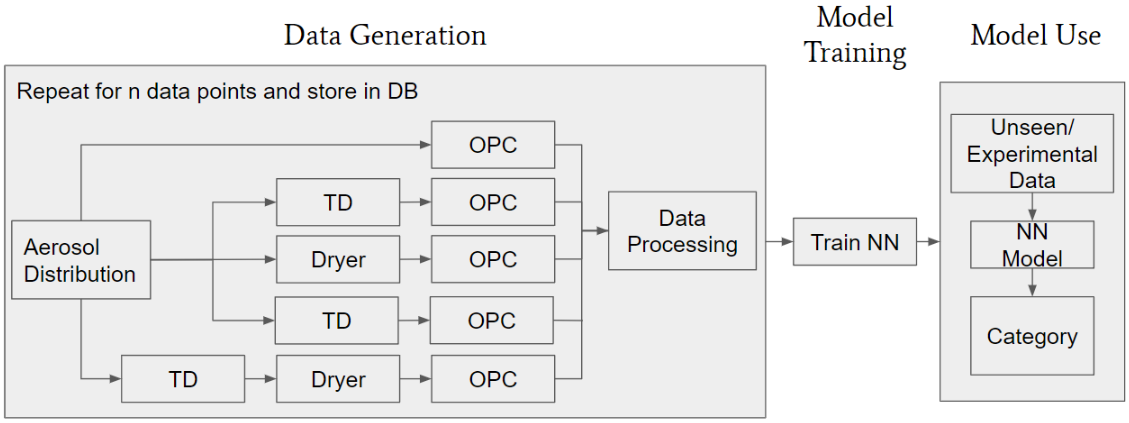

Figure 2-1: High level software architecture. The left and center sections of the image show how the model was constructed, while the right of the image shows how this trained model could be used in the future with experimental data

2.2 OPC Simulation

2.2.1 opcsim

Opcsim is a python library developed by David Hagan to simulate low-cost OPCs and Nephelometers [16]. This library was made to better quantify these instruments and their errors due to different conditions such as varied RH or mass loading. This is done using Mie theory with several assumptions to simplify the calculations. Opcsim allows one to create an instrument object (OPC or Nephelometer), and an Aerosol Distribution object then show the response of that instrument object to a given Aerosol Distribution at a certain ambient RH.

Both of these objects require a series of inputs. An OPC object requires a number of bins, angles of measurement (θ), laser wavelength (λ), and min and max detectable particle sizes. For this work the OPC object was set to match an Alphasense N3 OPC with 24 bins, θ from 32 to 98 degrees, λ of 639 nm, minimum particle diameter of 0.35 μm, and max particle diameter of 40 μm. These values were taken from the Alphasense website as well as from measurements provided by David Hagan. This OPC object then needs to be calibrated for a given refractive index, in this case $1.5 + 0j$ (from Alphasense website) [17].

Aerosol Distribution objects consist of one to many modes with each mode requiring the κ value from κ-Köhler theory (discussed more below), a density, the number of particles per cubic centimeter, the average diameter of those particles ($\overline{D_p}$), and the geometric standard deviation of the distribution (σ_g). The density, κ, and $\overline{D_p}$ values used were aerosol specific, and the number of particles in a given mode was normalized in later code so it was of minimal consequence. σ_g and $\overline{D_p}$ values were taken from representative aerosol distributions in Seinfeld and Pandis, 2016 [29]. One limitation of this approach is the availability of κ values, many aerosols have not had their κ values studied yet and thus could not be simulated in opcsim without modification to the code.

The code development for this paper was split into multiple phases, the first of which was the development of an experimental simulator which built upon opcsim. This simulator's main output is measurements from an OPC under specified experimental conditions which mirror the real output from the Alphasense N3 OPC. To do this the simulator modeled the effects of temperature and RH on the opcsim parameters discussed above.

2.2.2 Modeling Effects of RH

Using dryers in parallel with low-cost OPCs to modify RH has been tested once before in Chacón-Mateos et al 2022, which used a low-cost dryer to mitigate the effects of high RH [30]. They ran an experimental setup with two conditions and found a 59% and 64% reduction in $PM_{2.5}$ counts when using a dryer. There has also been substantial theoretical work done on quantifying the effects of RH on particle growth. Petters and Kreidenweis, 2007, created a model which allows for the representation of hygroscopicity with only a single variable κ, called κ-Köhler theory [31]. This theory initially allowed for modeling of only completely soluble or insoluble components, but was later extended to "sparingly soluble" particles [32]. Most κ values used in this work came from an online database maintained by M.D. Petters [33]. A previous application of this theory to low-cost OPCs showed that using κ values from literature provided reasonable accuracy for high RH conditions, although it noted in situ hygroscopicity values as being more effective [34]. In opcsim RH effects are simulated by taking in the "dry" $\overline{D_p}$ (D_D), the kappa value (κ_{eff}), and the RH and passing it in to Equation 2.1 to get the "wet" $\overline{D_p}$ (D_W). A_w is the water activity, which is equal to $\frac{RH}{100}$.

$$D_w = D_d * \sqrt[3]{1 + \frac{a_w}{1 - a_w}\kappa_{eff}} \tag{2.1}$$

An example diagram showing the effects of changing the RH value is shown in Figure 2-2. Each bar in the histogram corresponds to the amount of particles in each size bin. One can see that the distribution shifts right (towards larger particles) as humidity increases. These distributions are not entirely lognormal due to the calibration method of the OPC object used in opcsim.

2.2.3 Thermal Denuder Modeling

In addition to modeling the effects of RH on different components, this work needed a way to model aerosol vaporization as it went through a TD. One paper presented a modeling approach for experimental TD measurements to create a theoretical framework for volatility distributions of organic aerosols [35], this model provided some of the equations to make the TD simulator shown below. The remaining bulk of the equations used for the TD model came from a paper which presented a similar volatility model in the context of low volatility compounds using the Hertz-Knudsen equation [36].

Modeling of the thermal denuder was more intensive that modeling varied RH. First the evaporation flux is found using Equation 2.2 where J_e is evaporation flux, p^0 is the saturation vapor pressure at the TD temperature, k is Boltzmann's constant, T is the temperature inside the TD, m is the molecular mass, and γ is the evaporation constant, which is assumed to equal 1. This assumption means that the evaporation flux is at the theoretical maximum, so decreases in aerosol size in this work are an upper limit.

$$J_e = \frac{\gamma p^0}{\sqrt{2\pi mkT}} \tag{2.2}$$

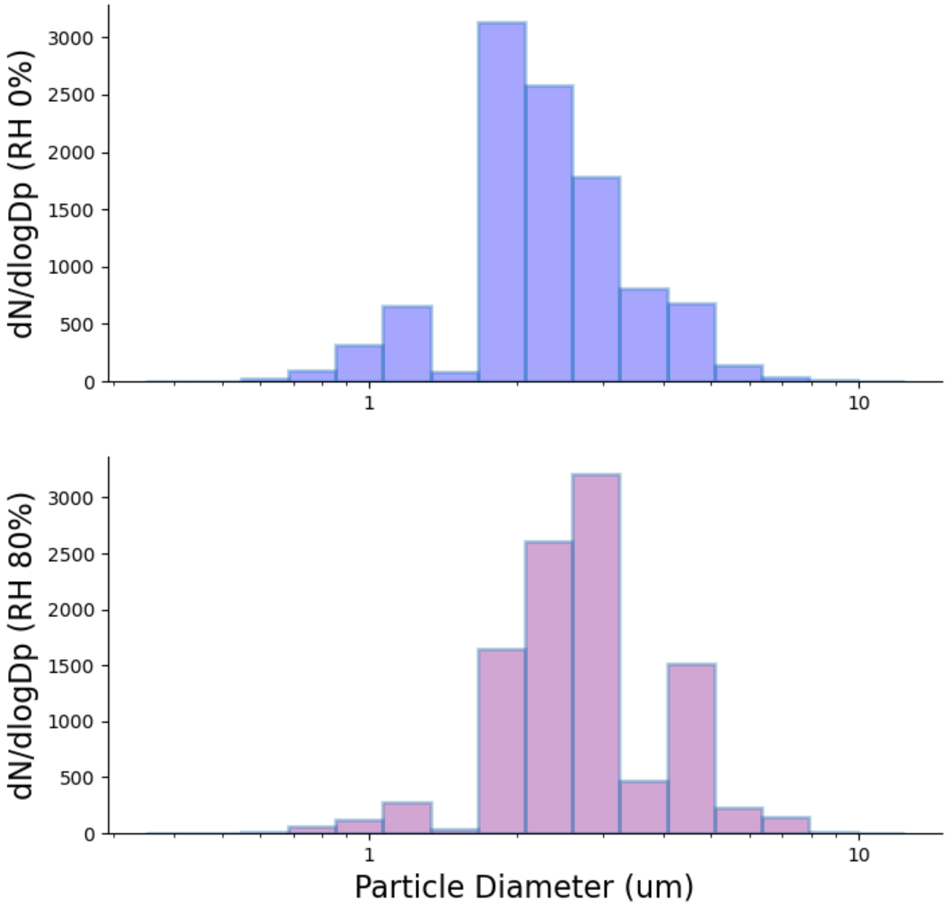

Figure 2-2: Example Histogram Output. The top figure shows an OPC output for a given aerosol at 0 RH, the bottom image shows the same aerosol at a higher RH. The OPC output shifts to the right (larger particles) with higher RH.

Saturation vapor pressure is calculated at the TD temperature T_{TD} using the saturation vapor pressure at a reference temperature p^0_{ref}, T_{ref}, the enthalpy of vaporization H_{vap}, the gas constant R and the Clausius-Clapeyron Equation 2.3.

$$p^0 = p^0_{ref} exp(\frac{\Delta H_{vap}}{R}(\frac{1}{T_{TD}} - \frac{1}{T_{ref}})) \tag{2.3}$$

Next the surface area (SA) of each particle is calculated with the assumption that the particles are spherical using Equation 2.4 where d_p is the particle diameter.

$$SA = \pi d_p^2 \tag{2.4}$$

Evaporation rate E can then be found by combining the evaporation flux and surface areas in Equation 2.5.

$$E = SAJ_e \tag{2.5}$$

The evaporation rate indicates how many molecules per second evaporate, to simplify calculations E was assumed to be constant for a small time step t_{step}, in this case chosen to be 0.01 seconds. One can then calculate the mass loss during that time step using Equation 2.6 and find the final mass after that time step using Equations 2.7 and 2.8.

$$\Delta mass = Et_{step}m \tag{2.6}$$

$$mass_i = \rho V \tag{2.7}$$

$$mass_f = mass_i = \Delta mass \tag{2.8}$$

If one then assumes the particle remains spherical and density remains constant with decreased mass, Equation 2.9 can be used to find the new radius after that time step.

$$V = \frac{m}{\rho} = \frac{4}{3}\pi r^3 \tag{2.9}$$

This process is then repeated for each successive time step until the total time is equal to the residence time of the TD (10 s). The output of running these equations will be a “thermally-denuded” particle diameter, which is then passed into opcsim as the $\overline{D_p}$ parameter along with the other required inputs discussed above.

There are several limitations to this approach. The first is that this requires additional data on physical properties of aerosols. Particularly the SVP is not available in literature for some compounds, which somewhat limits the aerosols this can be used on. A second limitation is that the flux equations used and the assumptions therein have only been experimentally verified for low-volatility aerosols and may be less effective for other compounds.

2.3 Data Generation

Each datapoint generated consisted of an OPCs response to five differing conditions with TD temperature and RH shown in Table 2.1. The first condition (298k, 20 RH) was meant to mimic ambient conditions while the next four were meant to be experimental conditions. What the ambient conditions are assumed to be is varied and discussed later to understand its impact. The experimental conditions were chosen to try to isolate the effects of temperature on aerosol volatility and RH on aerosol hygroscopicity. First the aerosol was simulated to go through the thermal denuder and then the effects of increasing RH were simulated, although the code was setup to support either order.

Each mode of each aerosol distribution would have an associated range of $\overline{D_p}$, σ_g, κ, molar mass, Δ_{hvap}, density, and number of particles. Using these properties an OPC output was generated to mimic the effects of temperature and RH on final particle diameter. Once this binned output was found, 10% randomness was introduced to try to avoid an overly idealized representation of a real OPC. This value is a hyperparameter of the model that could be further tuned once experimental results are obtained.

Eight categories of aerosols were chosen to be generated: four idealized "property-derived" categories, and four more realistic "compound-derived" categories. Properties for the eight categories are shown in Table 2.2. The property-derived aerosols were

Experimental Temperature (K)	Experimental Humidity
298	20
350	20
298	90
350	90
400	0

Table 2.1: Experimental Conditions. The first row corresponds to ambient conditions while the next four are experimental.

chosen to have saturation vapor pressure (SVP) and hygroscopicity (κ) values which were far apart from each other while keeping all other values the same. This was chosen so that the models could try to learn to differentiate the aerosols only using volatililty and hygroscopicity. The categories are NvHk (Nonvolatile High Kappa), VHk (Volatile High Kappa), NvLk (Nonvolatile Low Kappa), and VLk (Volatile Low Kappa). These are called "property-derived" since they have no direct relation to actual compounds and are used purely as a model testing tool.

The four compound-derived categories chosen were Urban Smog, Biomass Burning, Sea Salt Aerosol, and Dust. These categories had actual representative compounds chosen for them and real physical properties used for the different required values. Ammonium Sulfate was chosen to represent urban smog as it can be a major component of $PM_{2.5}$ in urban settings [37]. Levoglucosan was chosen to represent biomass burning aerosol. Levoglucosan is commonly associated with biomass burning aerosols and has been used as an indicator of biomass burning in the past [38] [39] [40]. Sodium Chloride was used to represent sea salt aerosol. Sea Salt Aerosol primarily consists of sodium chloride and it has been used in the past as a proxy for sea salt aerosol [41] [42]. The final proxy used is illite for dust aerosols. Illite has been previously used as a proxy for dust aerosols from clay, and contributes a large portion of mineral dust aerosol [43] [44]. The ideal formula of illite was used for the molar mass, density and kappa values. Illite is being assumed to be entirely nonvolatile at the temperatures considered in this work, which is supported by previous experimental work [45]. As such the saturation vapor pressure was set to 0 and the enthalpy of vaporization was given an arbitrary value.

Compound Name	SVP (Pa)	SVP Ref Temp(K)	Molar Mass (g/mol)	κ	ΔH_{vap} (kJ/mol)	Density (g/cm^3)
NvHk	0	405	150	1	225	1.2
NvLk	0	405	150	0.002	225	1.2
VHk	200	405	150	1	225	1.2
VLk	200	405	150	0.002	225	1.2
$(NH_4)_2SO_4$	3.75E-10	298	132.14	0.578	130	1.275
$C_6H_{10}O_5$	0.142	368	162.14	0.165	92.3	1.69
$NaCl$	0.615	1100	133.32	1.06	575	2.16
Illite	0	293	389.34	0.002	200	2.75

Table 2.2: Baseline Aerosol Properties, sources: [46] [31] [47] [48] [49] [50] [51] [52] [53] [54] [55]

The aerosol distribution properties are shown in Table 2.3. The property-derived category properties were all chosen to have the same ranges to try to maximize the difficulty of the learning problem and isolate volatility and hygroscopicity effects on the OPC outputs. The particle count choice was abstracted away by the data processing and was thus set arbitrarily to 1000 particles per cubic centimeter. The $\overline{D_p}$ range was chosen to be very wide to ensure that the features being learned depended on many different bin counts. The lower limit of 500 nanometers was set based off of the lower detection limit of the AlphaSense OPC keeping in mind that the accuracy is only 50% at the absolute lower limit. If the model was only trained on very small particles it might learn to ignore the larger bins and thus limit its applications. The σ_g range was chosen to encompass the values shown in Seinfeld and Pandis for most aerosol modes that a low-cost OPC could detect [29]. Additionally all of the parameters shown in Tables 2.2 and 2.3 were later varied to see the effect each parameter individually had on aerosol classification accuracy. This is discussed in depth below.

For the four compound-derived categories particle counts were adjusted based on relative diameters such that a roughly equal volume of each aerosol would be passed to the OPC for each category in two mode aerosol distributions. For single mode distributions all particle counts were set to 1000. σ_g ranges were again taken from Seinfeld and Pandis and set to be a conservatively wide range so as not to limit model applications. $\overline{D_p}$ ranges for each particle were also taken generally from Seinfeld and

Pandis. Each category of aerosol does not have a set $\overline{D_p}$ range but these values seek to roughly capture the relative size differences between the categories.

Category Name	Geometric Mean Diameter Range (μm)	Particle Count (#/cc)	σ_g Range (μm)
NvHk	0.5-10	1000	1.2-1.6
NvLk	0.5-10	1000	1.2-1.6
VHk	0.5-10	1000	1.2-1.6
VLk	0.5-10	1000	1.2-1.6
Urban Smog	0.1-0.5	10000	1.2-1.6
Biomass Burning	0.25-1	5000	1.2-1.6
Sea Salt	1-2.5	1000	1.2-1.6
Dust	1-10	500	1.2-1.6

Table 2.3: Baseline Aerosol Properties

Two types of aerosol distributions were generated: single mode and two mode. The single mode distributions consisted of one mode of one of the eight aerosol categories. This allowed for a four-category categorization problem for both the property and compound derived aerosol category sets. For two mode distributions two categories were picked (with replacement) from either set of 4. This led to the number of possible aerosol combinations expanding from 4 to 10 for each of the two aerosol category sets. Most of this work is concerned with two mode distributions as the problem of classification of single mode distributions turned out to be fairly trivial and less representative of real world conditions.

For single mode distributions 50-100 datapoints were generated from each of the eight categories. Each datapoint consists of an OPC's simulated response to the first four conditions in Table 2.1 for single mode distributions. This amount is relatively small, but was enough to perform classification perfectly on validation sets. For two mode distributions about 50-100 datapoints for each category were also initially generated using the properties shown in Table 2.3. After this initial data generation was performed it was determined that the sample size was too small for two mode classification to be performed adequately, so additional two mode aerosols were generated.

These additional two mode datapoints were generated in two different ways.

First aerosols with the baseline properties in Table 2.3 were generated. For the property-derived compounds, these were the only datapoints generated. For the property-derived compounds a second set of datapoints were generated which took the baseline properties and varied of them. Either the σ_g, the particle count, the particle mean diameter, the κ value, or the volatility would be adjusted while keeping the other conditions at baseline. The values these properties were varied to are shown in Table 2.4. Particle count were adjusted across all aerosol categories to 2000, 3000, 4000, 5000 or 6000 particles per cubic centimeter, but the data processing discussed below should negate the effects of changing the particle count and cause the classification to perform the same on all of these categories. The σ_g range was adjusted across all categories and set at either 0.25-0.75 μm, 0.5-1.5 μm, or 1-3 μm. When adjusting κ values of the κs of the low hygroscopicity categories (NvLk, VLk) were set at either 0.1, 0.2, or 0.4 and the κ values of the high hygroscopicity categories (NvHk, VHk) were set at either 0.7, 0.9 or 1.1. All combinations of these low and high κ values were taken, so in total 9 pairs were tested. $\overline{D_p}$ range was kept uniform across categories and set to either 0.25-1 μm, 0.5-2 μm, 1-4 μm, or 0.25-2.25 μm.

Experimental Property	Possible Values
σ_g Range (μm)	0.25-0.75, 0.5-1.5, 1-13
Mean Diameter Range (μm)	0.25-1, 0.5-2, 1-4, 0.25-2.25
Particle Count (#/cc)	2000, 3000, 4000, 5000, 6000
High κ	0.7, 0.9, 1.1
Low κ	0.1, 0.2, 0.4
Volatility	0.1, 1, 10, 100, 1000, 10000
Ambient RH Range	60-70, 70-80, 80-90
High RH Range	10-20, 20-30, 30-40
Ambient Temperature Range (K)	270-280, 280-290, 290-300
TD Temperature Range (K)	355-365, 370-380, 385-395

Table 2.4: Varied Experimental Properties

To determine the range for which to vary non-zero volatilies (VHk, VLk), some intuition was needed on reasonable ranges for this experiment. To do this a volatility parameter V was constructed by manipulating Equations 2.2-2.9. The final formula is shown in Equation 2.10. This parameter is intended to capture all properties of an aerosol that could affect the size change due to a TD and is unitless. In this work each volatility value shown can be associated with a corresponding SVP value.

$$V = \sqrt{m} * \frac{p^0}{\rho} * exp(\frac{-1000 * \Delta H_{vap}}{R} * (\frac{1}{T_{TD}} - \frac{1}{T_{ref}})) \tag{2.10}$$

The minimum volatility was determined by looking at the highest experimental temperature (400 K) and determining what volatility would correspond to almost no shrinkage at that temperature for the residence time of the TD (10 s). The maximum volatility was determined by looking at the ambient temperature and determining what volatility would correspond to complete evaporation after 10 minutes. This longer time was chosen since this aerosol would not be able to be sampled at all in theory if it vanished completely at room temperature immediately. The corresponding values of V were roughly 0.001 and 15,000 respectively. As such the volatility was chosen to be either 0.1, 1, 10, 100, 1000, or 10,000 to avoid those extreme situations. These volatility values correspond to SVP values from 0.0157-1570 Pa.

To determine the range of effective temperature and RH conditions that this model can correctly classify for, datapoints were generated with different ambient and experimental conditions. This is shown in the final four rows of Table 2.4. First RH was varied, instead of the low RH value being assumed as 20% it was set to a randomly chosen number between either 10-20%, 20-30% or 30-40% RH. High RH was also varied, instead of being assumed to be 90% it was set to a randomly chosen number between either 60-70%, 70-80%, or 80-90% RH. This gave nine different high and low RH combinations to test. Next, data was generated with varying high and low temperature values. Instead of ambient temperature being assumed at 298 K, it was randomly chosen from a range between either 270-280 K, 280-290 K, or 290-300 K. The high temperature was varied in a similar manner in ranges 355-365 K,370-380 K, and 385-395 K, leading to nine different temperature combinations to test.

Between all scenarios roughly 27,000 datapoints were generated in total for the two mode aerosol distributions with the four property-derived aerosols. Since there were many datapoints generated for the same set of somewhat narrowly restrictive conditions it was a concern that the model would simply memorize the datapoints, but the added 10% randomness worked to alleviate this potential problem. Data generation for the compound-derived categories was much simpler. Since the properties used had physical value, they were not varied and about 9,000 datapoints were generated with the properties referenced in tables 2.2 and 2.3 using the same 10% randomness.

2.4 Data Processing

Four data processing modes were created for the OPC outputs. These different modes were created with the intention of reducing the number of datapoints needed to successfully classify aerosols for the different distributions. The first and simplest was "Full Output" which takes the full time series OPC output from each of the experimental conditions and concatenates them together in a dataframe. A five OPC setup (one for each condition) where each OPC has 24 bins would produce 120 bin counts per time step. The simulated experiments used were generally 10 minutes long for each datapoint and the sampling rate of the OPCs was 5 seconds, meaning that if you kept every bin count from every OPC the datapoints became very large. As such this data processing mode was not used in the results. But this mode could prove useful for those trying to generate realistic synthetic OPC outputs.

The second data processing mode developed was "Time Averaged", which took the average bin counts for each experiment, normalized them to abstract away particle count and output those 24 bin counts for each OPC. For the same experiment discussed above this would reduce the parameters per datapoint to 120 ($24 * 5$) regardless of experiment length or OPC sampling rate. This reduced datapoint size allowed this mode to be much more practical, although it does require a good amount of training examples for use with a neural network. This mode has the advantage of not using any human selected features.

The third data processing mode used was "Median Diameter". This was the mode that reduced the data down to the fewest number of parameters per datapoint. First the average particle size of the aerosol distribution at ambient conditions (20 RH, 298 K) was found by taking the weighted average of the different bins and their median diameters. For each subsequent experimental condition there were four parameters of the aerosol distribution: the average size change between the baseline diameter and the experimental diameter, the experimental RH, the experimental temperature, and the absolute average particle size. The size change parameter was set to 1 if the experimental aerosol distribution was fully evaporated and the baseline aerosol distribution was not. This allowed a representation of an aerosol distribution with four experimental conditions with only 16 parameters. This mode was very efficacious for single mode aerosols and allowed accurate classification using minimal datapoints, but was ineffective for two mode distributions. This was likely due to the fact that all of the parameters generated related to the aerosol distribution as a whole and did not allow for any finer grain data to be extracted.

For two mode aerosol distributions the "Multi Component" mode was developed. To do this the assumption was made that each mode in an aerosol distribution would be roughly Gaussian relative to the bin counts and dN/dlogDp axes. While this is not a perfect assumption, based on observation modes did tend to be roughly noisy normal distributions. With this assumption in place a two mode Gaussian distribution was fit to the time averaged OPC output for each experimental condition using a Gaussian Mixture Model. The function used to fit this distribution also indicated if the output was only a single mode. This essentially allowed the splitting of the two modes for a given aerosol. Using this information some features were selected: the difference between the median diameter of each mode, the total OPC count across all bins for each mode, if there was a second mode, the absolute difference between the maximum value of each mode, and the difference in the weights between the two modes. First this information was captured for the ambient conditions (20 RH, 298K), then for each of the four experimental conditions those parameters were captured and made relative to the values in the ambient conditions. This was done to capture changes in size from baseline rather than absolute particle size. This allowed for representation of a datapoint with 20 parameters (5 for each experimental condition). When data available was on the order of hundreds to a few thousand datapoints this was the most effective data processing mode for two mode aerosols. Upon generation of further datapoints the time averaged mode began to perform better.

2.5 Aerosol Classification

Once data was generated and processed classification of the aerosols could begin. The tool initially used for single mode four-category classification was AutoML from the mljar-supervised package [27]. This package generates and tests Linear, LightGBM, Decision Tree, XgBoost, Neural Network and Random Forest algorithms automatically with just the input of data and its associated class. It split the data into train and validation sets, then output statistics from these models such as a confusion matrix, the learning curve, the precision recall curve, feature importance and the ROC curve. For the single mode distributions AutoML was able to perform very well with very few lines of code as well as provide clear visualizations of what was happening. For single mode distribution almost all of these algorithms tested were able to provide near zero loss on the validation set. But when this model was tested on an unseen test set it did not perform as well, thus a simple fully connected neural network was used.

For two mode aerosol distributions AutoML was also used initially for the ten class classification problems. This performed relatively well, although not as well as the single mode, on the validation sets. Particularly with the Multi-Component mode AutoML seemed able to classify well with only a few hundred datapoints for each category. But after the model created for two mode classification was optimized it was ran on unseen test data, and the accuracy was substantially lower than that seen on the validation data. This indicates that the model was over fitting. Additionally AutoML does have specific guidelines for the models that it generates and optimizing them within these fairly restrictive parameters could force the model to miss out on optimal solutions. Due to this apparent over fitting the AutoML results were simply used as a guideline for creation of a custom model.

Figure 2-3 shows the results of the various models that AutoML ran on the two mode classification problem with the y axis being a loss metric, in this case the logloss, and the x-axis being the type of model tested. Each model type is shown as a box and

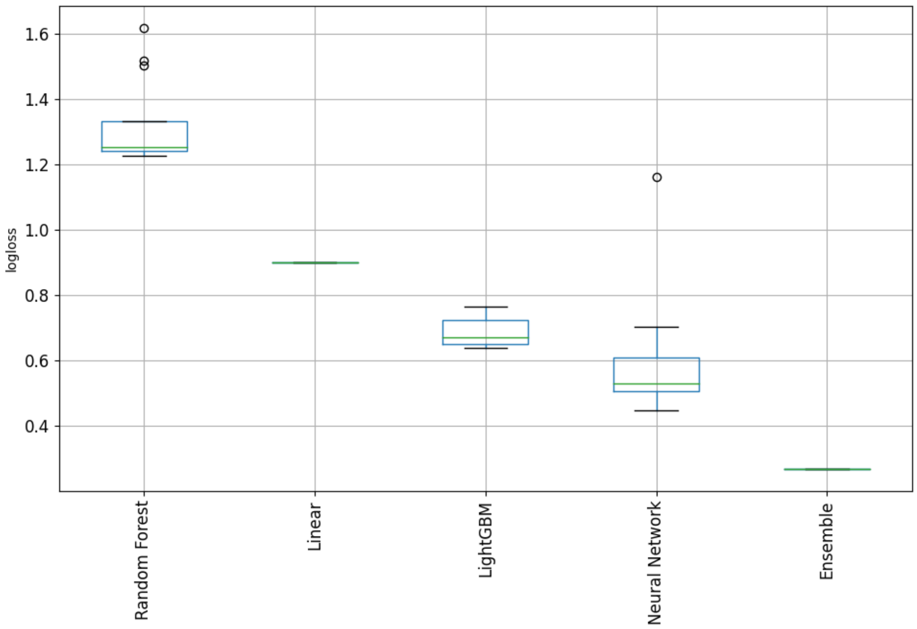

Figure 2-3: AutoML Model Performance Box Plot. This shows the example loss for an AutoML output with a box plot for the different models tested in the same category.

whisker plot since AutoML tested multiple different variations for each type. This plot was generated from data using the "Time Averaged" processing method, and similar figures were generated for the other data processing methods. The ensemble method is a combination of the other four types of models, and thus cannot be recreated directly without repeating the over fitting. This figure suggests that the best class of model for this problem is a neural network. As such a fully connected neural network (NN) was generated using the Keras package. A fully connected neural network was chosen versus other neural network architectures due to the relatively small number of parameters.

The NN used has 2 hidden dense layers with ReLU activation functions and 94 and 64 neurons per hidden layer respectively. The final dense layer has 10 neurons and a SoftMax activation function applied. The input shape depends on the data processing method used for the input data. The learning rate was set to 0.001. These hyperparameters for the model were optimized using the KerasTuner library, which allows for a grid search between many hyperparameter options [28]. This optimization was performed using validation loss after a 80:10:10 train, test, validation split. The optimization ran 50 hyperparameter combinations with 5 trials per combination and the above architecture was found to perform amongst the best combinations. After the model architecture had been determined it was trained for up to 500 epochs with an early stopping callback in place to prevent over fitting. This same NN architecture was used on both the property and compound derived categories.

The effects of ignoring each of the five experimental conditions was also studied on classification behavior. Since this method is meant to be low-cost and deployable, the need for fewer OPCs, dryers/humidifiers, and simple TDs is very desirable. Determining the number of apparatuses necessary for a certain model performance is therefore necessary to aid the future construction of the array that is being simulated in this work. Each experimental condition output consisted of a row in a Pandas DataFrame for every datapoint, one row at a time per point was excluded then the NN was retrained and validated with this data. This was intended to show the relative need of each experiment, which could be particularly relevant to any field work that might occur in the future for an array such as this.

2.6 Experimental Laboratory Setup

One challenge of this work is that the results presented are solely from simulation. While these results are promising, if a setup like the one described here were to be constructed it would need real world data to prove and calibrate the simulation. To begin verifying the simulation work preliminary laboratory work has been performed. Although results have not been entirely analyzed and collected yet, a working setup has been constructed for potential future use. An overview of this setup is shown in Figure 2-4. Aerosols are first generated using a TSI Aerosol Generator 3076, and then pushed through into an ARI TD. This atomizer can generate particles from 0.01-2 μm and the TD is capable of generating temperatures up to 200 C. This TD also has a bypass flow which allows the aerosol to go through unperturbed if ambient temperature is desired.

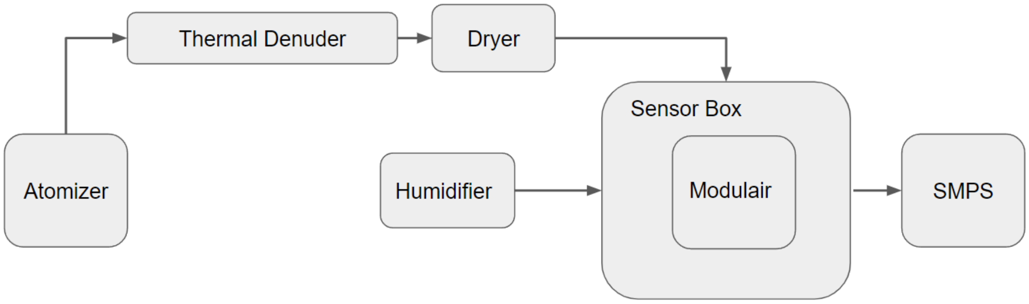

Figure 2-4: Experimental Setup. Aerosols are generated from an atomizer and then have RH and temperature modulated before being sent in to a sensor box.

After passing through the TD the RH of the flow is modulated. If low RH flow is desired, a Nafion dryer is supplied clean air to dry the flow. If high RH is desired the flow passes through the Nafion unperturbed and a humidifier feeds high RH air into the sensor box. This setup was able to achieve RH values between 10-70%. After passing through the dryer, aerosols are then fed into a sensor box which contains a QuantAQ Modulair. The Modulair is a commercial low-cost air quality sensing device which includes an RH and temperature sensor as well as an AlphaSense N3 OPC. The Modulair samples from the air in the sensor box once a minute. In addition to the Modulair, a Scanning Mobility Particle Sizer (SMPS) sampled the air in the sensor box. The SMPS is a much higher rated instrument and provides very accurate particle counts to compare the OPC to. Some limitations of this setup are that it can only generate aerosols for a small part of the OPC detection range, and that RH can not reach the 90% value that was used in the simulation work. But the particle sizes it can generate are those of greatest interest since many of the anthropogenic sources of PM produce aerosols in this size range, and these smaller sizes can be associated with worse health outcomes [56]. The RH values generated represent a slightly more difficult problem than simulated since it should be easier to estimate if an aerosol is hygroscopic at a higher RH. Thus if experimental results could verify the efficacy of this simulation approach for these more difficult conditions it would be particularly valuable.

2.7 Methodology Summary

In conclusion there are five main parts to this simulation and experimental work. The first and smallest part is integration with the existing code from opcsim. The second part is development of a simulator for a thermal denuder, which has been based around the Hertz-Knudsen equation. The third section is the data processing, this section goes through the different data processing modes one by one. It is important to note that most of the results presented below come from the Time Averaged processing mode, but that mode only became the best choice after a large number of datapoints were generated. Thus the decision for someone seeking to produce something similar may be different depending on availability of this data. Fourth the classification algorithm used for all of the results below is presented. The choice of a neural network was based on AutoML testing multiple model configurations and the hyperparameters of the model were set via a hyperparameter tuning library. Finally a laboratory setup was proposed which could produce conditions near those simulated, and was able to create particles in the size range of particular interest for human health. This could be used in future experimental work to close the sim-to-real gap between that may exist with the proposed simulator.

THIS PAGE INTENTIONALLY LEFT BLANK

Chapter 3

Aerosol Classification Results

3.1 Single Mode Classification

3.1.1 Property-Derived Categories

The first and simplest case that results were obtained for was the classification of single mode aerosol distributions for the four different property-derived aerosols. For the single mode classification 100 simulated experiments with each aerosol were generated, the Median Diameter mode was used for data processing, and AutoML was used to generate the classification algorithm. AutoML validation results were fairly far apart in accuracy (difference greater than 30%) from the test results. Even with AutoML being unable to provide the desired test results it was still able to provide useful comparisons between model architectures as well as feature importance metrics which validated the choice of the Median Diameter data processing mode. These insights led to the use of the 2 layer fully connected neural network architecture discussed in the methodology section with the Median Diameter processing. This allowed the model to correctly identify the incoming aerosol for every simulated experiment. The normalized test confusion matrix is shown in Figure 3-1. This figure shows the model's prediction on the X axis and the true aerosol on the Y axis, with a result on the diagonal indicating a correct classification.

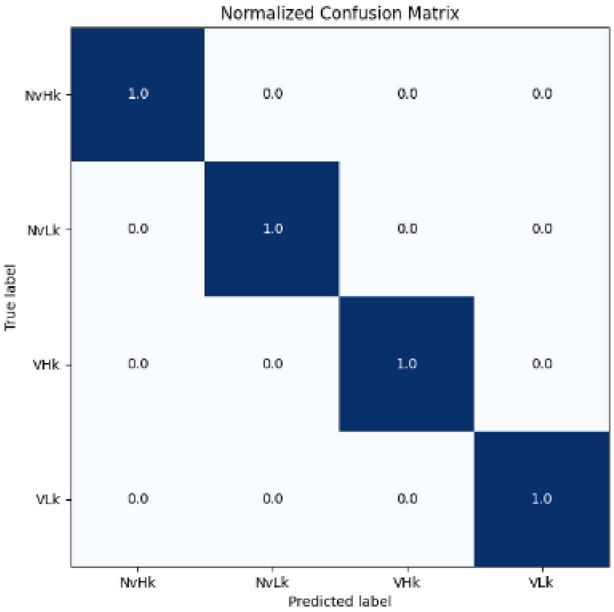

Figure 3-1: Property-Derived Single Compound Confusion Matrix. The model is able to perfectly classify single mode distributions.

3.1.2 Compound-Derived Categories

Once proof of concept had been established this same method of data generation and processing was then used on the compound-derived aerosols. The results of retraining the same neural network with datapoints generated from the compound-derived aerosols is shown in Figure 3-2. Again AutoML was able to perfectly classify every datapoint in the validation set, but on the test set the accuracy was substantially lower. Using the NN ameliorated these problems and led to perfect test accuracy. These results as well as those for the single mode property-derived aerosols were generated with only three of the four experimental conditions, leaving out the 400 K, 0 RH condition. This suggests that if single mode classification were the only objective, three experimental conditions would be all that was necessary. This result, or at least the necessary temperature/RH in the experimental conditions, might change were one to use different proxies for each category.

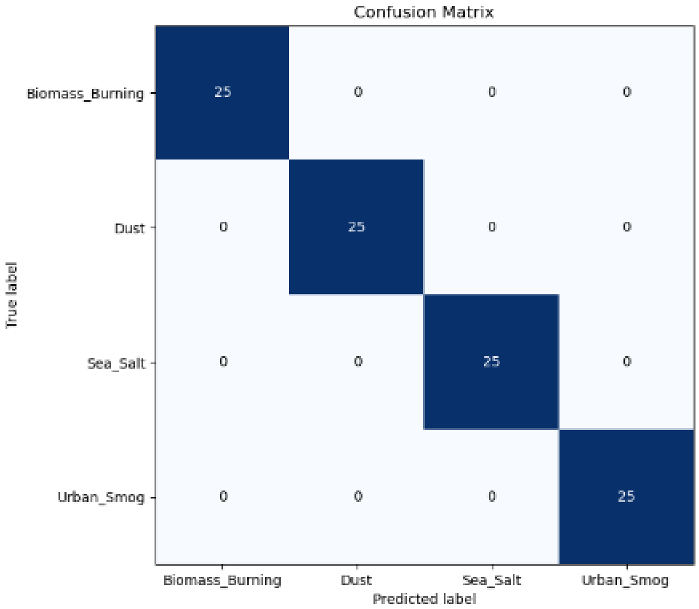

Figure 3-2: Compound-Derived Single Compound Confusion Matrix. The model is able to perfectly classify single mode distributions.

3.2 Two Mode Aerosol Distribution Classification

3.2.1 Property-Derived Categories

A more difficult and more realistic question to approach is resolving each part of a two mode aerosol distribution. This is a ten-category classification problem instead of four and many more datapoints were required for this problem to get accurate results than the single mode distributions. These datapoints were split 80:10:10 between the train, validation, and test sets then fed into a NN. This NN was able to achieve 94.9% accuracy on the test set and the performance broken down by category is shown in Figure 3-3. Abbreviations used are explained in Table A.1. This clearly demonstrates proof of concept that two mode distributions can have both of their modes identified under idealized conditions. Additionally while data availability may be a limitation for this approach generally, the decoupling of the categories of interest from exact compounds somewhat ameliorates this problem. If one knows generally comparative values for the hygroscopicity and volatility of a category that could be good enough for

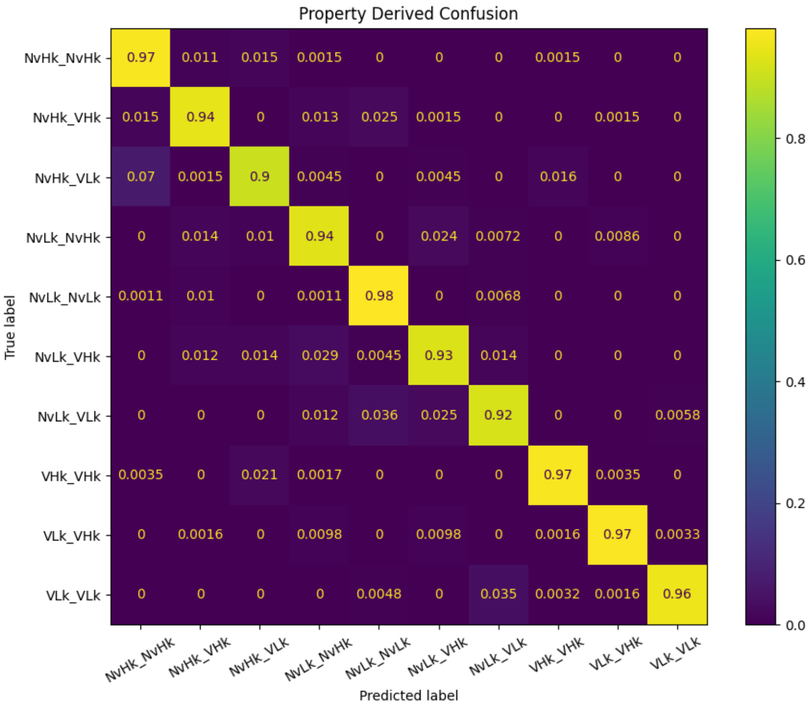

Figure 3-3: Property-Derived Two Mode Confusion Matrix. The model is able to categorize both elements in a two mode aerosol distribution well, with relatively evenly distributed confusion. Abbreviations given in Table A.1.

usable results. This overall accuracy includes data with both the baseline properties shown in Table 2.2 as well as all datapoints generated under the varied conditions described in the model sensitivity section. A further breakdown of those results is shown later.

3.2.2 Compound-Derived Categories

This same method of data processing was used on two mode aerosol distributions composed of the four compound-derived categories. A separate NN with the same architecture as above was trained on this data with an 80:10:10 train, validation, test split. The test accuracy was 99.3%, and the performance broken down by category is shown in Figure 3-4 where dust is shown as D, biomass burning is shown as BB, urban smog is shown as S, and sea salt it shown as SS. For example D_SS is a two

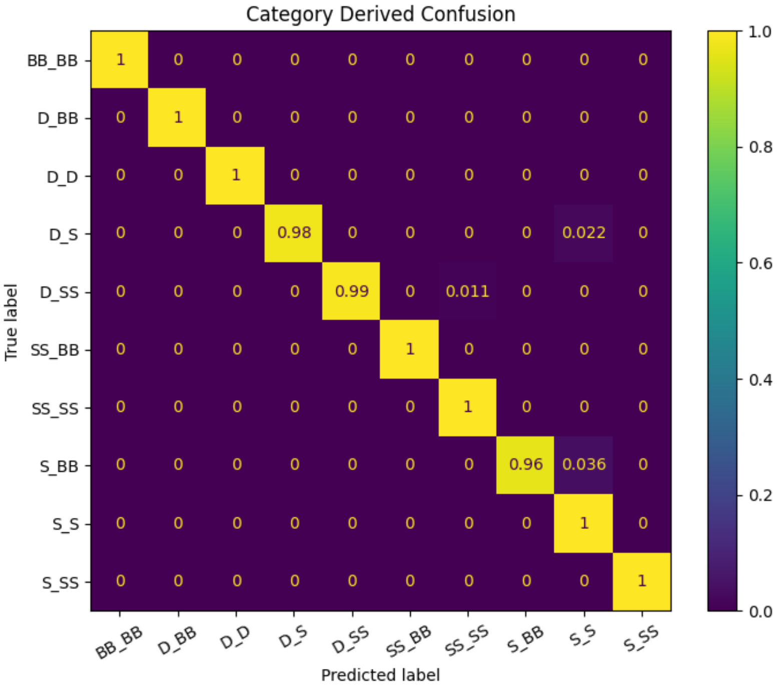

Figure 3-4: Compound-Derived Two Mode Confusion Matrix showing a fairly uniform very high performing model. This indicates that under constant conditions the NN can achieve near-perfect classification. Abbreviations shown in Table A.1.

mode aerosol distribution containing one mode of dust aerosol and another of sea salt aerosol. These results again show the efficacy of using this NN approach, and help tether this approach to real world conditions. The accuracy of this model is likely higher than the property-derived categories because conditions were not varied in this data set, unlike the property-derived data set.

3.3 Model Sensitivity

3.3.1 Experimental Condition Occlusion

A NN with the same architecture as discussed was trained with only four of the five experimental conditions shown in Table 2.1 for each datapoint, to test the dependence of the two mode results on each experimental condition for the property-derived categories. Table 3.1 shows the full occlusion results. Leaving out the 350 K 20 RH, 298 K 90 RH, or 350 K 90 RH conditions all led to a decrease in classification accuracy of less than 3% from the baseline (94.9%), while dropping the 400 K 0 RH, or 298 K 20 RH condition led to a slightly larger drop in accuracy. This suggests that if a real world experimental array were to be constructed one could simplify the construction and reduce the cost by leaving out a condition.

Extending on this, two of the five conditions at a time were occluded from each datapoint. The biggest drop in accuracy was seen when leaving out both of the conditions at ambient temperature. This makes sense as it gets rid of an important comparison point to measure volatility from. The most notable results of this condition occlusion were that two of the experiments with two conditions occluded were still able to achieve a test accuracy of 90% or higher. Both of these experiments correspond to leaving out one of the ambient temperature and one of the high temperature conditions. The experiment leaving out the 350 K 20 RH, and 350 K 90 RH conditions is particularly interesting when considering practical experimental construction. This would only leave the baseline condition as well as one high RH and one high temperature condition and has accuracy within a percent of the baseline value. An experimental array for these conditions would need just one humidifier, and one simple TD.

Experimental Apparatus Occlusion

Each low-cost apparatus included in the experimental array would be associated with additional logistical challenges were this array to be produced. As such the results of only modulating RH or temperature, and therefore only investigating either hygroscopicity or volatility, was explored. Using only the conditions at ambient temperature the model accuracy drops to 36.8%, and using only the conditions at ambient RH the model accuracy is 75.9%. Based on laboratory work performed, RH is assumed to remain unchanged after passing through the TD. It should be noted that both these results do depend on the ambient RH and temperature assumption made. These results indicate that the inclusion of both apparatuses substantially outperforms one or the other, and that the use of a TD is absolutely crucial.

Condition(s) Occluded	Classification Accuracy
298 K 20 RH	89.8%
350 K 20 RH	93.3%
298 K 90 RH	92.7%
350 K 90 RH	93.7%
400 K 0 RH	86%
298 K 20 RH, 350 K 20 RH	85.5%
298 K 20 RH, 298 K 90 RH	72.0%
298 K 20 RH, 350 K 90 RH	88.8%
298 K 20 RH, 400 K 0 RH	80.0%
350 K 20 RH, 298 K 90 RH	91.4%
350 K 20 RH, 350 K 90 RH	93.9%
350 K 20 RH, 400 K 0 RH	84.7%
298 K 90 RH, 400 K 0 RH	84.2%
298 K 90 RH, 350 K 90 RH	87.5%
400 K 0 RH, 350 K 90 RH	77.3%

Table 3.1: Experimental Conditions Occlusion results showing the effects of leaving out 1-2 of the conditions in each datapoint.

3.3.2 Effects of Varying Parameters

As described in the methodology section, the data for the property-derived aerosols was generated with each parameter varied in multiple discrete ranges of values. As a result of this, the performance for each discrete set of each parameter can be analyzed to get a better idea of the model's effective range and limitations for each parameter. A test data set of about 5000 two mode datapoints was used, so that there would be sufficient examples of each condition for each parameter, and none of the datapoints would have been seen during NN training. Each category discussed below is the result of keeping all baseline properties of the aerosols, shown in Tables 2.2 and 2.3, the same besides the parameter being discussed. The same NN trained on all of the property-derived datapoints in section 3.2.1 was used to test all datapoints below. An overview of the results is shown in Table 3.2.

Baseline Conditions, Particle Counts, and Mean Diameter

The results of this analysis showed that for the baseline conditions, aerosols with varying geometric mean diameter, and aerosols varying particle counts, the model

Varied Parameter	Classification Accuracy Range	Reasons for Decreased Performance
Particle Count	100%	None
Mean Diameter Range	100%	None
κ	95-100%	Decreased gap between low and high κ values
σ_g Range	84-100%	Shrinking σ_g
Volatility	36-100%	Very low and very high volatility values

Table 3.2: Classification accuracy of trained model on aerosols with one conditions changed from baseline. The trend causing decreased performance is shown for parameters which didn't have 100% accuracy for all conditions.

was able to correctly classify every sample. The baseline results being perfect is very promising and also makes sense as those were the conditions most commonly seen in training of the NN. The results of the compound-derived categories, which did not vary parameters, were also near perfect. The effect of particle counts was abstracted away by normalizing OPC bin counts so the largest was equal to 1. Therefore it makes sense that those results would be the same as the baseline results. For the differing geometric mean diameter ranges, these results are very promising. The model performing as well on aerosol distributions centered around sub-micron particle sizes as it did on 2-4 μm geometric mean diameters is a great result and shows the use of setting such a wide particle size range for the baseline conditions. Particles below 1 μm are frequently of particular interest to research applications and this lack of sensitivity to size shows this could be a potential tool for that research. While this result is promising it should be noted that previous experimental work has shown a drop in accuracy below 0.8 μm, so the conditions simulated may be overly idealized at this particle size [20].

Varied κ

The datapoints generated from augmenting the κ values also had promising results. The model was able to classify with above 95% accuracy for every combination of low and high κ values tested. This included a combination where the high κ values were

0.7 and the low κ values were 0.4, which was perfectly classified. This suggests that the κ parameter is able to be very accurately estimated using this approach and can be used for more than just very hygroscopic and entirely non-hygroscopic aerosols.

Varied σ_g

The two properties which had the largest effects on classification accuracy were the σ_g range and the volatility. Each aerosol distribution had a σ_g randomly set to a value within a specified range. Lowering the values in that range had an adverse effect on accuracy. The 1-3 μm σ_g range was able to be classified perfectly, while the 0.5-1.5 μm σ_g range had a classification accuracy of 92%, and the 0.25-0.75 μm σ_g range had a classification accuracy of 84%. This follows logically, particularly for the smaller particle sizes. If a distribution has a much smaller variance and a geometric mean diameter of 0.5 μm it will be captured only by the one or two smallest sized bins for the OPC tested. This could make a decrease in size harder to identify since the distribution would begin in the smallest bin. A larger σ_g range allows for a more Gaussian distribution amongst the bins, where one might expect it to be easier to track the size changes due to hygroscopicity and volatility. The same effect would be seen at particle sizes approaching the upper size detection limit of the OPC. The takeaway from these results is that a wider distribution is easier to classify, and one should be particularly careful with narrow distributions around the edges of the detection range of an OPC.

Varied Volatility

Varying volatility of the volatile compounds had the largest effect on classification accuracy amongst all of the properties. These categories involved varying the volatility of the two high volatility categories (VLk and VHk) while keeping the nonvolatile categories entirely nonvolatile. To find the range of volatilities relevant to this work, Equation 2.10 was used. The lowest volatility that could be captured by this setup would be something that has barely any size loss at the highest TD temperature tested (400 K). To find this volatility, the baseline properties of the property-derived aerosols were passed into the volatility equation and the SVP was iteratively lowered until 10 seconds at 400 K for a 0.5 μm particle produced a 1% decrease in mean diameter. This resulted in V $\approx$ 0.001, where V is the parameter to describe the volatility of an aerosol shown in Equation 2.10. The highest volatility of relevance was set to be an aerosol which would shrink to 0 after 10 minutes at ambient temperature (298 K). This resulted in V $\approx$ 15,000. Since these very high volatility aerosols would almost never be seen by an OPC sampling ambient air, and a 1% decrease in size could not reasonably be seen with these instruments, volatilities in the range of 0.1-10000 were tested (corresponding to changing SVP from 0.0157-1570 Pa while keeping other conditions at baseline). Results for this are shown in Figure 3-5.

For volatility in range $V = 10 - 1000$ (corresponding to an SVP of 1.57 Pa - 157 Pa at 400 K) the model was able to perfectly classify incoming aerosols. At $V = 1$ (corresponding to an SVP of 0.157 Pa at 400 K) the model performance began to drop to around 95%, and at $V = 0.1$ (corresponding to an SVP of 0.0157 Pa at 400 K) the model performance crashed to 36%. At very high volatility $V = 10000$ (1570 Pa at 400 K) the accuracy also began to drop slightly to 96%. These results suggest that with the given experimental temperatures there is a fairly well defined range of volatility values that can be classified against nonvolitile compounds with the same hygroscopicity.

Decreasing the volatility values of the two volatile categories is essentially making them more and more similar to their corresponding nonvolatile counterpart. It makes sense that when volatility is set low enough such that only the 400 K conditions can produce any size change the accuracy tends to drop off. A value of $V = 0.1$ corresponds to a small size decrease from a 400 K TD with a 10 second residence time, and no size decrease at 350 K. A higher max temperature would likely be needed to classify aerosols with volatility in that range vs nonvolatile aerosols with the same hygroscopicity. A $V = 10000$ volatility value corresponds to an aerosol which is only able to last less than an hour at room temperature before evaporating completely for the particle diameters of interest to this work. In practice an aerosol with this level of volatility would very rarely be detected by an instrument monitoring ambient

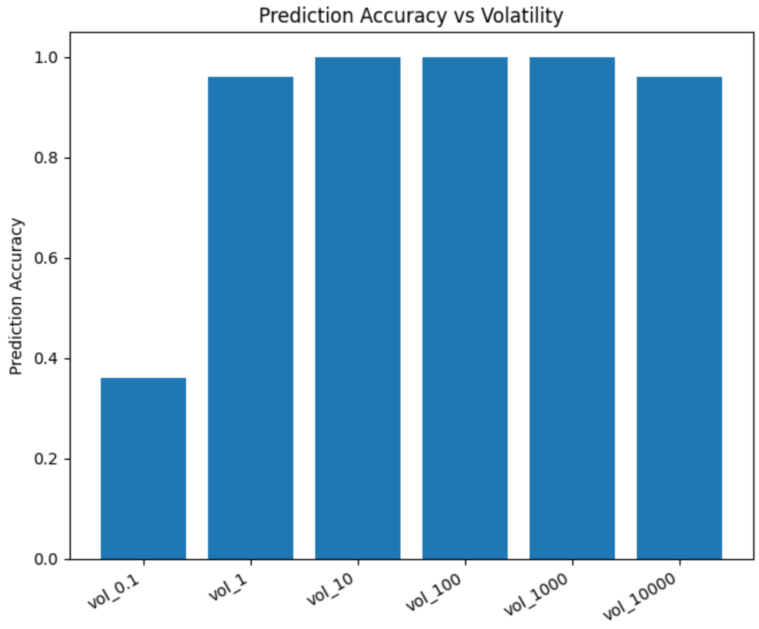

Figure 3-5: Results of varying high volatility SVP while keeping nonvolatile aerosols at SVP=0. Model accuracy drops off when the volatility of two otherwise identical aerosols become very close (SVP=0 vs SVP=0.0157). The model performs above 95% for all higher volatilities.

PM unless the OPC was very near the PM source. Any volatility above this value would be almost entirely impractical for this application, so its ability to classify that volatility above 95% is promising. This small inaccuracy is likely due to obfuscation of the two volatile compounds since both VLk and VHk would disappear for all heated conditions.

3.3.3 Varied RH and Temperature

Next, instead of varying the properties of the aerosols themselves, the properties of the experimental conditions shown in Table 2.1 were varied for the two mode property-derived aerosol distributions and tested against the NN trained on all data. Similar to the last section, new datapoints were generated under these new conditions to ensure the NN had not already seen any of the test data.

Table 3.3 shows the effects of varying the high and low RH conditions. For these results the RH for the 400 K experimental condition was set to the ambient RH range, while it was set to 0 for the initial model. This was done after preliminary lab results suggested that a 400 K temperature did not lead to a large change in RH. This was potentially due to the air lowering back towards ambient temperature at the time of sampling. Table 3.3 shows that moving away from the entirely idealized RHs produces a drop in classification accuracy of 7-13%. For the lab setup proposed in the methodology section, the theoretical accuracy would still be around 90%. These results are encouraging for future work. To test these setups in a laboratory setting a relative humidity value of 90% is fairly hard to achieve, and these accuracies suggest that it might not be necessary for good results. For the aerosols used to produce these results the κ values of the non-hygroscopic and hygroscopic aerosols were very far apart. Were these values closer together the results of varying RH conditions may become more of a concern.

Ambient RH Range	High RH Range	Classification Accuracy
10-20	60-70	86.3%
10-20	70-80	89.5%
10-20	80-90	89.3%
20-30	60-70	90.2%
20-30	70-80	90.9%
20-30	80-90	91.9%
30-40	60-70	92.7%
30-40	70-80	92.2%
30-40	80-90	92.9%

Table 3.3: Varied experimental condition results showing the effect of changing ambient and high RH.

Table 3.4 shows the result of changing the TD and ambient temperatures. These results also have a slight change from the trained NN, in that these results were created with only a ambient and high temperature instead of two different high temperatures. This was again chosen with practical concerns in mind, as only requiring one high temperature could simplify TD construction and possibly reduce the number of TDs needed. The condition occlusion results suggested that performance drops from using only one TD were negligible, so these results could help inform future experimental work more directly.

The consistency of the accuracy stands out immediately. Every temperature combination scored within a percent of every other combination. This suggests that for these aerosol volatilities, any high TD temperature between 355-395 K should work well for any ambient temperature between 270-300 K. These results would likely change with the volatility of the high and low volatility compounds. The results from the varied volatility section should provide some insight on how best to select max TD temperature given the volatility of the compounds of interest. Plugging in an aerosol's properties to the volatility formula would provide information on what temperatures are required for a substantial change in the diameter of that aerosol.

Ambient TD Temperature Range (K)	High TD Temperature Range (K)	Classification Accuracy
270-280	355-365	93.3%
270-280	370-380	93.2%
270-280	385-395	92.8%
280-290	355-365	92.6%
280-290	370-380	93.1%
280-290	385-395	93.3%
290-300	355-365	93.1%
290-300	370-380	93.0%
290-300	385-395	93.2%

Table 3.4: Varied TD condition results showing the effect of changing ambient and TD temperature

3.4 Summary of Results

These results show strong proof of concept that (under idealized conditions) an OPC in concert with other simple apparatuses can determine basic compostional information of ambient aerosols. These results held for both a single mode four-category problem as well as a two mode ten-category problem. Additionally these results proved to be robust to changing input conditions. The model performed very well for a reasonable range of volatilities, with the accuracy only waning at volatilites corresponding to a very minimal change at the highest TD temperature tested. These results were also somewhat robust to occluding experimental conditions. This could allow one interested in producing an experimental array to simplify their task and reduce their cost with minimal loss of accuracy. Finally it was shown that in there is some robustness to what the ambient temperatures and RHs are assumed to be.

3.5 Limitations

There are some limitations to the application of these results to keep in mind. Most of the robustness studies were performed on aerosols with κ values which were very far apart, and aerosols which were either relatively volatile or entirely nonvolatile. Future work should be done to test the robustness to changing conditions with compound-derived aerosols which may be more similar to each other than the the property-derived aerosols used are. Additionally future work should be done to test these results in both laboratory and real-world settings.

Chapter 4

Conclusion

Knowing the composition and sources of $PM_{2.5}$ is an important and difficult problem. With source data, regulation of PM as well as further research on its negative health effects are made much easier. The existing solutions are expensive and have poor spatial resolution, but produce very high quality data. The motivation of this work is that currently no lower cost option for a more basic level of information exists. This lower cost could allow for monitors capturing compositional PM data to be spread more broadly and have high spatiotemporal resolution or simply reduce the barrier to entry for a single point of composition data.

This work shows that it is possible to get a level of compositonal data using low-cost apparatuses, which should lead to easier source attribution. A very simple fully connected neural network was able to perfectly classify single mode aerosol distributions for both property and category derived compounds, and was able to classify two mode aerosol distributions with an accuracy of greater than 90%. These distributions do a fair job of approximating real world conditions, and importantly the model is robust to different aerosol properties. This holds even between aerosols which could both be considered moderately hygroscopic or moderately volatile. Since the categories chosen for this work were broad, it was important to demonstrate this robustness to a wide range of initial conditions and chosen compound values.

Aerosols were able to be distinguished from one another even when all properties besides hygroscopicity and volatility measures were held constant. This suggests that those two parameters alone are enough to identify differences in aerosols to a fair degree. This extreme condensing of the scope of features allowed for learning to be done even with noisy data, and a relatively small number of datapoints. These two properties also have the advantage of being able to be investigated fairly easily using low-cost, small apparatuses. Taken as a whole, this approach takes away many of the needs that necessitate the very complex and expensive instrumentation that is used in traditional approaches for determining particulate matter composition. In addition to the simulation work, a lab setup was proposed to begin to test the real world applicability of this work.

This work has a few limitations to note. The first, and most likely largest, is that this work get results purely from simulation. The sim-to-real gap could be substantial and is currently unquantified. Software development was done such that adjustments could be made easily, but this would be necessary future work if real world experiments were to be carried out. Initial experiments could use the laboratory setup proposed in the methodology section. These results would enable the amount of noise in the simulation to be tuned, and would highlight areas where the simulation results and experimental results were most different. If the simulation was doing a poor job of classifying certain volatilities, Equations 2.2-2.9 could be augmented or added to. If the simulation was doing a poor job of classifying certain hygroscopicities, a correction factor could be added to the κ values or opcsim could be augmented to support other metrics of hygroscopity.

A second limitation is the number of parameters required for each compound in an aerosol distribution. Data availability, particularly for the hygroscopicity, proved somewhat of a limitation when choosing representative compounds for each property-derived category. Thirdly the generalizability of the equations used for volatility is something which warrants more study, and would again need to be further verified by experimental work.

Partially stemming from these limitations, there is substantial future work enabled by the novel approach presented. This software could be fairly easily tested with real world inputs, first in a lab setting and later in the field. Software was developed to determine the similarity between an experimental result with known conditions and a simulated result under those same conditions. As a first step each of the four compound-derived aerosols could be tested under the conditions described using the proposed lab setup. This would present a single mode problem that the model should perfectly classify. If errors were present it would become apparent and the simulated results could be compared to the real world results to see what was being incorrectly simulated. Based on these errors it would be possible to identify where the model should be further tuned and test the assumptions inherit in the simulator. Testing if density does remain constant with decreased mass is of particular interest.

If an experimental array were constructed, these arrays could be a low-cost, deployable solution investiage the composition of PM and could be easily dispersed and moved. Further development of the software could also be undertaken to extend this approach to three mode aerosol distributions, which are frequently seen in the real world. This could also potentially work to close the sim-to-real gap.

The main significance of this work is helping to lower the barrier to entry for compositional data/source attribution of ambient PM. This work provides proof of concept that data beyond binned PM counts can be determined from low-cost OPCs. This enables future work to collect real world compositional data with these apparatuses. Even if this attribution is fairly basic, this approach is temporally and spatially high resolution and could provide at least some data to regions which may not have any. Policy makers could have a better idea of what to focus regulation on rather than blindly trying to reduce PM as a whole, citizen scientists could potentially use this data to determine local PM sources, and researchers could use this data to help ameliorate the data shortage problems that many studies suffer from when trying to link PM source and composition to negative health outcomes.

THIS PAGE INTENTIONALLY LEFT BLANK

Appendix A

Tables

Abbreviation	Full Category Name
NvHk	Nonvolatile High Kappa
NvLk	Nonvolatile Low Kappa
VHk	Volatile High Kappa
VLk	Volatilie Low Kappa
S	Urban Smog
SS	Sea Salt
BB	Biomass Burning
D	Dust

Table A.1: Abbreviations used for aerosol categories.

THIS PAGE INTENTIONALLY LEFT BLANK

Bibliography

[1] Guanghe Wang et al. "Effects of ozone and fine particulate matter (PM2. 5) on rat system inflammation and cardiac function". In: *Toxicology letters* 217.1 (2013), pp. 23–33.

[2] Shaolong Feng et al. "The health effects of ambient PM2. 5 and potential mechanisms". In: *Ecotoxicology and environmental safety* 128 (2016), pp. 67–74.

[3] Richard Burnett et al. "Global estimates of mortality associated with long-term exposure to outdoor fine particulate matter". In: *Proceedings of the National Academy of Sciences* 115.38 (2018), pp. 9592–9597.

[4] Arideep Mukherjee and Madhoolika Agrawal. "World air particulate matter: sources, distribution and health effects". In: *Environmental chemistry letters* 15 (2017), pp. 283–309.

[5] Philip M Fine, Constantinos Sioutas, and Paul A Solomon. "Secondary particulate matter in the United States: insights from the particulate matter supersites program and related studies". In: *Journal of the Air & Waste Management Association* 58.2 (2008), pp. 234–253.

[6] CA Belis et al. "Critical review and meta-analysis of ambient particulate matter source apportionment using receptor models in Europe". In: *Atmospheric Environment* 69 (2013), pp. 94–108.

[7] Federico Karagulian et al. "Contributions to cities' ambient particulate matter (PM): A systematic review of local source contributions at global level". In: *Atmospheric environment* 120 (2015), pp. 475–483.

[8] Philip K Hopke et al. "Global review of recent source apportionments for airborne particulate matter". In: *Science of The Total Environment* 740 (2020), p. 140091.

[9] World Health Organization et al. *Review of evidence on health aspects of air pollution: REVIHAAP project: technical report.* Tech. rep. World Health Organization. Regional Office for Europe, 2021.

[10] Sverre Vedal et al. "National Particle Component Toxicity (NPACT) initiative report on cardiovascular effects." In: *Research Report (Health Effects Institute)* 178 (2013), pp. 5–8.

[11] Kate Adams et al. "Particulate matter components, sources, and health: Systematic approaches to testing effects". In: *Journal of the Air & Waste Management Association* 65.5 (2015), pp. 544–558.

[12] WM Vatauk, WL Klotz, and RL Stallings. *EPA Air Pollution Control Cost Manual*. Tech. rep. EPA 452/B-02-001, 1999.

[13] Paul A Solomon et al. "US national PM2. 5 chemical speciation monitoring networks—CSN and IMPROVE: description of networks". In: *Journal of the Air & Waste Management Association* 64.12 (2014), pp. 1410–1438.

[14] Nga L Ng et al. "An Aerosol Chemical Speciation Monitor (ACSM) for routine monitoring of the composition and mass concentrations of ambient aerosol". In: *Aerosol Science and Technology* 45.7 (2011), pp. 780–794.

[15] Gustav Mie. "Beiträge zur Optik trüber Medien, speziell kolloidaler Metallösungen". In: *Annalen der physik* 330.3 (1908), pp. 377–445.

[16] David H Hagan and Jesse H Kroll. "Assessing the accuracy of low-cost optical particle sensors using a physics-based approach". In: *Atmospheric measurement techniques* 13.11 (2020), pp. 6343–6355.

[17] *OPC-N3 OPC-N3 particle particle monitor monitor - alphasense.com.* URL: `https://www.alphasense.com/wp-content/uploads/2022/09/Alphasense_OPC-N3_datasheet.pdf`.

[18] Andrea Di Antonio et al. "Developing a relative humidity correction for low-cost sensors measuring ambient particulate matter". In: *Sensors* 18.9 (2018), p. 2790.

[19] David H Hagan et al. "Inferring aerosol sources from low-cost air quality sensor measurements: a case study in Delhi, India". In: *Environmental Science & Technology Letters* 6.8 (2019), pp. 467–472.

[20] Sinan Sousan et al. "Evaluation of the Alphasense optical particle counter (OPC-N2) and the Grimm portable aerosol spectrometer (PAS-1.108)". In: *Aerosol Science and Technology* 50.12 (2016), pp. 1352–1365.

[21] Leigh R Crilley et al. "Evaluation of a low-cost optical particle counter (Alphasense OPC-N2) for ambient air monitoring". In: *Atmospheric Measurement Techniques* 11.2 (2018), pp. 709–720.

[22] Jose M Barcelo-Ordinas et al. "Calibrating low-cost air quality sensors using multiple arrays of sensors". In: *2018 IEEE Wireless Communications and Networking Conference (WCNC)*. IEEE. 2018, pp. 1–6.

[23] Roberto Gasparini, Runjun Li, and Don R Collins. "Integration of size distributions and size-resolved hygroscopicity measured during the Houston Supersite for compositional categorization of the aerosol". In: *Atmospheric Environment* 38.20 (2004), pp. 3285–3303.

[24] AE Faulhaber et al. "Characterization of a thermodenuder-particle beam mass spectrometer system for the study of organic aerosol volatility and composition". In: *Atmospheric Measurement Techniques* 2.1 (2009), pp. 15–31.

[25] VA Lanz et al. "Characterization of aerosol chemical composition with aerosol mass spectrometry in Central Europe: an overview". In: *Atmospheric Chemistry and Physics* 10.21 (2010), pp. 10453–10471.

[26] Costa D Christopoulos et al. “A machine learning approach to aerosol classification for single-particle mass spectrometry”. In: *Atmospheric Measurement Techniques* 11.10 (2018), pp. 5687–5699.

[27] Aleksandra Płońska and Piotr Płoński. *MLJAR: State-of-the-art Automated Machine Learning Framework for Tabular Data. Version 0.10.3.* Łapy, Poland, 2021. URL: `https://github.com/mljar/mljar-supervised`.

[28] François Chollet et al. *Keras.* `https://keras.io`. 2015.

[29] John H Seinfeld and Spyros N Pandis. *Atmospheric chemistry and physics: from air pollution to climate change.* John Wiley & Sons, 2016.

[30] Miriam Chacón-Mateos et al. “Evaluation of a low-cost dryer for a low-cost optical particle counter”. In: *Atmospheric Measurement Techniques* 15.24 (2022), pp. 7395–7410.

[31] MD Petters and SM Kreidenweis. “A single parameter representation of hygroscopic growth and cloud condensation nucleus activity”. In: *Atmospheric Chemistry and Physics* 7.8 (2007), pp. 1961–1971.

[32] MD Petters and SM Kreidenweis. “A single parameter representation of hygroscopic growth and cloud condensation nucleus activity–Part 2: Including solubility”. In: *Atmospheric Chemistry and Physics* 8.20 (2008), pp. 6273–6279.

[33] M.D Petter. *Kappa Database.* `https://mdpetters.github.io/kappa/`. 2023.

[34] Leigh R Crilley et al. “Effect of aerosol composition on the performance of low-cost optical particle counter correction factors”. In: *Atmospheric Measurement Techniques* 13.3 (2020), pp. 1181–1193.

[35] E Karnezi, Ilona Riipinen, and SN Pandis. “Measuring the atmospheric organic aerosol volatility distribution: a theoretical analysis”. In: *Atmospheric Measurement Techniques* 7.9 (2014), pp. 2953–2965.

[36] Christopher D Cappa, Edward R Lovejoy, and AR Ravishankara. “Determination of evaporation rates and vapor pressures of very low volatility compounds: a study of the C4- C10 and C12 dicarboxylic acids”. In: *The Journal of Physical Chemistry A* 111.16 (2007), pp. 3099–3109.

[37] Boming Ye et al. “Concentration and chemical composition of PM2. 5 in Shanghai for a 1-year period”. In: *Atmospheric Environment* 37.4 (2003), pp. 499–510.

[38] Man Nin Chan et al. “Hygroscopicity of water-soluble organic compounds in atmospheric aerosols: Amino acids and biomass burning derived organic species”. In: *Environmental science & technology* 39.6 (2005), pp. 1555–1562.

[39] Vladimir O Elias et al. “Evaluating levoglucosan as an indicator of biomass burning in Carajas, Amazonia: A comparison to the charcoal record”. In: *Geochimica et Cosmochimica Acta* 65.2 (2001), pp. 267–272.

[40] Hemraj Bhattarai et al. “Levoglucosan as a tracer of biomass burning: Recent progress and perspectives”. In: *Atmospheric Research* 220 (2019), pp. 20–33.

[41] CA Randles, LM Russell, and V Ramaswamy. "Hygroscopic and optical properties of organic sea salt aerosol and consequences for climate forcing". In: *Geophysical Research Letters* 31.16 (2004).

[42] E Schindelholz, BE Risteen, and RG Kelly. "Effect of relative humidity on corrosion of steel under sea salt aerosol proxies: I. NaCl". In: *Journal of The Electrochemical Society* 161.10 (2014), p. C450.

[43] Nadine Hoffmann et al. "Contact freezing efficiency of mineral dust aerosols studied in an electrodynamic balance: quantitative size and temperature dependence for illite particles". In: *Faraday discussions* 165 (2013), pp. 383–390.

[44] Daniel B Curtis et al. "A laboratory investigation of light scattering from representative components of mineral dust aerosol at a wavelength of 550 nm". In: *Journal of Geophysical Research: Atmospheres* 113.D8 (2008).

[45] Srivats Srinivasachar et al. "Mineral behavior during coal combustion 2. Illite transformations". In: *Progress in Energy and Combustion Science* 16.4 (1990), pp. 293–302.

[46] Gennady J Kabo et al. "Experimental and theoretical study of thermodynamic properties of levoglucosan". In: *The Journal of Chemical Thermodynamics* 85 (2015), pp. 101–110.

[47] Curtis T Ewing and Kurt H Stern. "Equilibrium vaporization rates and vapor pressures of solid and liquid sodium chloride, potassium chloride, potassium bromide, cesium iodide, and lithium fluoride". In: *The Journal of Physical Chemistry* 78.20 (1974), pp. 1998–2005.

[48] *Ammonium sulfate.* 2023. URL: `https://www.chemspider.com/Chemical-Structure.22944.html`.

[49] *Calcium carbonate.* URL: `http://www.chemspider.com/Chemical-Structure.9708.html`.

[50] ChemSpider. *Levoglucosan.* 2023. URL: `https://www.chemspider.com/Chemical-Structure.9587432.html`.

[51] Markus D Petters et al. "Chemical aging and the hydrophobic-to-hydrophilic conversion of carbonaceous aerosol". In: *Geophysical research letters* 33.24 (2006).

[52] James Chia-san Chou. "Thermodynamic properties of aqueous sodium chloride solutions from 32 to 350 F". PhD thesis. 1968.

[53] Ryan C Sullivan et al. "Timescale for hygroscopic conversion of calcite mineral particles through heterogeneous reaction with nitric acid". In: *Physical Chemistry Chemical Physics* 11.36 (2009), pp. 7826–7837.

[54] Hanna Herich et al. "Water uptake of clay and desert dust aerosol particles at sub-and supersaturated water vapor conditions". In: *Physical Chemistry Chemical Physics* 11.36 (2009), pp. 7804–7809.

[55] Dave Barthelmy. URL: `http://www.webmineral.com/data/Illite.shtml#.ZCC0inbMKUm`.

[56] Athanasios Valavanidis, Konstantinos Fiotakis, and Thomais Vlachogianni. "Airborne particulate matter and human health: toxicological assessment and importance of size and composition of particles for oxidative damage and carcinogenic mechanisms". In: *Journal of Environmental Science and Health, Part C* 26.4 (2008), pp. 339–362.