doi.org/10.26434/chemrxiv.9177146.v1

Excitation Emission Matrix Fluorescence Spectroscopy for Combustion Generated Particulate Matter Source Identification

Jay Rutherford, Neal Dawson-Elli, Anne M. Manicone, Gregory V. Korshin, Igor V. Novosselov, Edmund Seto, Jonathan D. Posner

Submitted date: 20/08/2019 • Posted date: 21/08/2019

Citation information: Rutherford, Jay; Dawson-Elli, Neal; Manicone, Anne M.; Korshin, Gregory V.; Novosselov, Igor V.; Seto, Edmund; et al. (2019): Excitation Emission Matrix Fluorescence Spectroscopy for Combustion Generated Particulate Matter Source Identification. ChemRxiv. Preprint.

The inhalation of particulate matter (PM) is a significant health risk associated with reduced life expectancy due to increased cardio-pulmonary disease and exacerbation of respiratory diseases such as asthma and pneumonia. PM originates from natural and anthropogenic sources including combustion engines, cigarettes, agricultural burning, and forest fires. Identifying the source of PM can inform effective mitigation strategies and policies, but this is difficult to do using current techniques. Here we present a method for identifying PM source using excitation emission matrix (EEM) fluorescence spectroscopy and a machine learning algorithm. We collected combustion generated PM2.5 from wood burning, diesel exhaust, and cigarettes using filters. Filters were weighted to determine mass concentration followed by extraction into cyclohexane and analysis by EEM fluorescence spectroscopy. Spectra obtained from each source were used as machine learning training data for source identification in mixed samples. This method can predict the presence or absence of the three laboratory sources with an overall accuracy of 89% when the threshold for classifying a source as present is 1.1 µg/m3 in air over a 24-hour sampling time. We apply this method to a small set of field samples to evaluate its effectiveness.

Excitation Emission Matrix Fluorescence Spectroscopy for Combustion Generated Particulate Matter Source Identification

Jay W. Rutherford,[1] Neal Dawson-Elli,[1] Anne. M. Manicone,[5] Gregory V. Korshin,[2] Igor V. Novosselov,[2] Edmund Seto,[4] Jonathan D. Posner [*1,2,3]*

[1]Department of Chemical Engineering, [2]Department of Mechanical Engineering, [3]Department of Family Medicine, [4]Environmental and Occupational Health Sciences, [5]Department of Medicine: Pulmonary, Critical Care and Sleep Medicine University of Washington.

Abstract

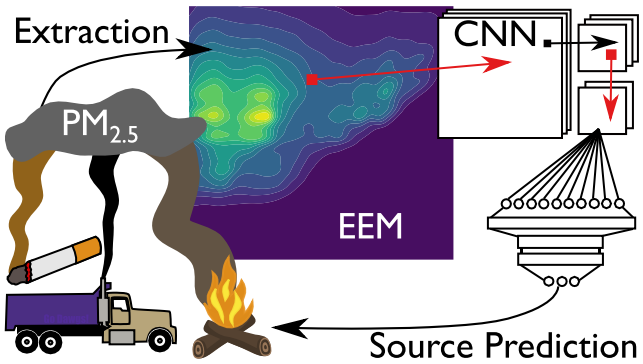
The inhalation of particulate matter (PM) is a significant health risk associated with reduced life expectancy due to increased cardio-pulmonary disease and exacerbation of respiratory diseases such as asthma and pneumonia. PM originates from natural and anthropogenic sources including combustion engines, cigarettes, agricultural burning, and forest fires. Identifying the source of PM can inform effective mitigation strategies and policies, but this is difficult to do using current techniques. Here we present a method for identifying PM source using excitation emission matrix (EEM) fluorescence spectroscopy and a machine learning algorithm. We collected combustion generated $PM_{2.5}$ from wood burning, diesel exhaust, and cigarettes using filters. Filters were weighted to determine mass concentration followed by extraction into cyclohexane and analysis by EEM fluorescence spectroscopy. Spectra obtained from each source were used as machine learning training data for source identification in mixed samples. This method can predict the presence or absence of the three laboratory sources with an overall accuracy of 89% when

the threshold for classifying a source as present is 1.1 µg/m^3 in air over a 24-hour sampling time. We apply this method to a small set of field samples to evaluate its effectiveness.

Keywords

Fluorescence, Source Apportionment, Particulate Matter, Diesel, Woodsmoke, Neural Network

Graphical Abstract



1. Introduction

According to the Global Burden of Disease study, air pollution is the world's largest environmental health risk accounting for 4.9 million deaths and 147 million disability adjusted life years in 2017.[1] Particulate matter (PM) originates from a wide range of natural and anthropogenic sources. Most PM is a result of emissions directly from sources such as diesel engines, agricultural burning, cooking with biomass, electrical power generation, pollen, bacteria, and soil. In addition to these primary sources, PM results from chemical processes in the atmosphere, referred to as secondary sources.[2] The human respiratory system is effective at removing many of the particles that enter the respiratory tract before they enter deep into the lungs. Larger particles are removed by impaction and sedimentation in the upper respiratory tract and branching airways of the lungs while smaller particles can penetrate deeper into the lungs.[3] A size cut-off of 2.5 µm in aerodynamic equivalent diameter ($PM_{2.5}$) and smaller has been established for the purpose of regulation of PM pollution.[4] The USEPA standards for maximum exposure levels are 12 µg/m^3 annual average and 35 µg/m^3 daily average.[5] The World Health Organization (WHO) guidelines are lower at 10

and 25 μg/m^3, respectively.[6] Studies suggest that some sources of PM are worse for health than others. For example, black carbon, which is associated with traffic, has been shown to be worse for health than $PM_{2.5}$ alone,[7,8] but the body of evidence as a whole does not conclusively show what sources or combinations thereof are the worst for health.[9–12]

Asthma is known to be exacerbated by $PM_{2.5}$ exposure.[13] In practice, clinicians advise asthmatic patients to avoid exposure to pollution that they are sensitive to and to avoid exertion outdoors when air quality is poor.[14] A study using parental questionnaires and proximity to roadways found increased asthma risk in children with exposure to second hand smoke, but not with roadway proximity.[15] User-friendly and inexpensive tools for monitoring source specific PM exposure will enable continued and more quantitative epidemiological research in the area of source specific health impacts and may enable regulations targeting the worst sources of PM pollution.[16]

In this work, we use excitation-emission matrix (EEM) fluorescent spectroscopy and machine learning to identify the source of PM. Fluorescence spectroscopy is a sensitive analytical technique with the ability to detect fluorescence from a single molecule using sophisticated instrumentation.[17] With widely available benchtop fluorimeters limits of detection are around 1 ng/mL for polycyclic aromatic hydrocarbons, a common chemical component of PM air pollution.[18,19] For this reason, fluorescence spectroscopy is an attractive analytical technique for PM analysis due to typical sample sizes of PM being small. Although fluorescence is a very sensitive technique it is not highly specific due to many analytes having overlapping spectra. EEM spectroscopy can increase the specificity of fluorescence spectroscopy by collecting fluorescent emission spectra at multiple excitation wavelengths, giving a 2D dataset or matrix of fluorescence intensities.[20] EEM spectroscopy has been widely applied to analysis of complex environmental water samples[21] as well as analysis of atmospheric PM.[22–28] Mladenov et. al. suggested EEM could be useful as a source identification tool for atmospheric PM, but did not evaluate the ability of EEM alone to identify sources.[26] Other work applying EEM to atmospheric aerosols discusses the chemical composition of various regions of fluorescence, but does not discuss using EEM as a source apportionment or identification tool.[22–25,27,28]

EEMs provide complex spectral information consisting of thousands of wavelength dependent fluorescent intensities (~20,000 datapoints for the EEMs in this work), as such, a variety of approaches have been used to interpret EEM spectra. Fluorescent regional integration considers specific regions of an EEM spectrum based on compounds of interest that display fluorescence in various regions (e.g. aromatic protein and humic acid like regions when applied to water samples).[23] This approach has the advantage of simplicity, but it is unable to distinguish overlapping spectra which was a challenge in our samples. Techniques used for interpreting EEM spectra that can handle overlapping spectra include partial least squares regression (PLS), parallel factor analysis (PARAFAC), and multivariate curve resolution (MCR). These techniques have been used successfully to identify specific chemical components of atmospheric PM, but have not been used for source apportionment.[18,22,25] In this work we used a convolutional neural network (CNN), a machine-learning technique that is well suited to handle 2D data like EEM spectra. Machine learning techniques, like a CNN, do not rely on underlying theory or assumptions to create a model, instead these techniques use data inputs and expected outputs, referred to as training data, to create a non-parametric model. The model framework is defined and then the training data are used to iteratively adjust model parameters to best fit the training data, this is the learning or training process.

A CNN is a combination of a neural network[29] preceded by a feature recognition process called convolution.[30] Individual steps in the network are called layers. The first layers are convolutional layers consisting of filters of user-defined size which are iteratively scanned, or convolved, across the input data. The filter values are randomly initialized and adjusted to identify relevant features during the training process. CNNs excel at processing 2D data (e.g. image classification[31,32]) due to their ability to learn patterns or features encoded in spatial data using the convolutional layers. The result of the convolution process is fed into a neural network to map the learned patterns and features to an output value or classification.

Neural networks and CNNs have been applied to a variety of PM exposure related problems. For example, a CNN has been used to predict continuous $PM_{2.5}$ concentrations from discrete measurements.[33] Source apportionment has been conducted using a neural network with elemental composition as the input,

which is similar to our work, but the input data was 1D elemental data so convolution layers are not used.[34] Neural networks have been used to process EEM spectra for analysis of water samples, contaminates in olive oil, and antibiotics in urine.[35–38] In these examples, neural networks are used because of their ability to fit non-linear behavior without the need for assumptions about the underlying data and their ability to interpret information from the entire EEM, both of which are applicable to our work.

In this paper, we demonstrate the ability of EEM coupled with a convolutional neural network to identify PM from woodsmoke, cigarettes and diesel exhaust with a limit of detection of 2.6 $\mu g/m^3$ based on a 24-hour sampling time. $PM_{2.5}$ samples were collected on PTFE filters using personal sampling devices. Samples were weighed to determine $PM_{2.5}$ mass concentration and then extracted in cyclohexane. The cyclohexane extracts were analyzed using EEM fluorescence. The spectra from the three sources show unique but overlapping fluorescent EEM spectra. We applied a CNN to identify the presence or absence of the three sources present in a set of EEM spectra consisting of one, two, or all three sources. We achieved an overall accuracy of 89% in identifying sources. This technique has a limit of detection well below the USEPA and WHO recommended exposure levels for PM and may be useful for personal monitoring in epidemiological studies of respiratory diseases such as asthma.

2. Materials and Methods

2.1 Particulate Matter Sampling and Extraction

We sampled $PM_{2.5}$ from cigarettes, diesel exhaust, and woodsmoke using 2.0 µm pore PTFE membrane filters (Pall Zefluor®, Pall Cat. # P5PJ037) housed in Harvard School of Public Health Personal Exposure Monitor (BGI, Butler, NJ Cat. # HP2518) sampling cassettes. Filters were operated at a flowrate of 1.8 lpm using either portable or stationary vacuum pumps (AirChek XR5000 pump, SKC Inc., Eighty Four, PA or VP0625-V1014-D2-0511, Medo USA, Roselle, IL with custom manifold of nine VFB-65-BV roto-meters, Dwyer Instrument, Michigan City, IN). Flow rates were verified using an air flow calibrator (Gilian Gilibrator PN# 800268, Sensidyne, St. Petersburg, FL).

We collected woodsmoke by burning 1 ½ by ¾ inch Douglas fir sticks cut from dimensional lumber in a prototype side-feed, natural-draft, improved cookstove. Our sampling devices were placed in a sealed chamber connected to the exhaust hood duct at the sampling point described for gravimetric sampling by Sullivan et. al.[39] We collected diesel exhaust particulate from the exposure room in UW's controlled inhalation diesel exhaust exposure facility.[40] The diesel PM is generated by a 435 cc direct injection single cylinder diesel engine (Yanmar LW Series) fueled with ultra-low-sulfur diesel. We collected cigarette smoke either by lighting cigarettes in a fume hood and allowing them to smolder or from the exposure chamber of a cigarette smoking machine (Model TE-10B, Teague Enterprises, Woodland, CA). The TE-10B produces mainstream smoke mixed with sidestream smoke from filtered 3R4F research cigarettes (Tobacco Research Institute, University of Kentucky, Lexington, KY). Two cigarettes were puffed simultaneously for 2 seconds for a total of 8 puffs, at a flow rate of 1.05 l/min. The smoke collected represents approximately 10% mainstream and 90% sidestream to more closely resemble secondhand smoke.

Following collection, filters are removed from the samplers and placed in a chamber with 37% (SD = 4%) relative humidity for 24 hours.[41] The filters were then weighed using a micro-balance with 0.5 µg resolution (Mettler-Toledo UMT-2, Greifensee, Switzerland). Initial weights of each filter are recorded in the same manner and we use the difference to calculate the mass of $PM_{2.5}$ collected.

We placed the filters into 20 mL glass vials (Cat # 89096-774 VWR, Edison, NJ), submerged the filters in cyclohexane (Uvasol® Cyclohexane for Spectroscopy, MilliporeSigma Cat. #1.02822.2500), and sonicated for 30 minutes (42 kHz, 2510R-MT Branson, Ultrasonic Corp., Danbury, CT). Filters were generally submerged in ~10 mL of cyclohexane to achieve an initial extract concentration of 5 µg PM/mL cyclohexane or greater. For filters with low PM loading we cut the filter into fourths to enable extraction in as little as 3 mL of cyclohexane to maintain extract concentrations at or above 5 µg/mL. Typically, the PM was not dislodged from the filter during extraction allowing for direct analysis of the extract. If significant PM was dislodged and suspended causing turbidity, the extract was filtered with a 0.2 µm PTFE syringe filter (VWR Cat. #28145-491) before analysis.

We collected a total of 37 filter samples and used the extracts from these filters to create 113 samples for EEM analysis. They consisted of 81 single source samples diluted to concentrations between 0.2 µg/mL and 10 µg/mL, 21 mixtures of the single source samples and five samples from filters with mixed PM from serial sampling of the sources. We also collected six spectra from liquid extracts of filters that were loaded into sampling devices and weighed, but no air was drawn through the filters (method blanks). For training our algorithm, we used 12 of the 113 EEM samples leaving a total of 101 samples for testing the algorithm. Table 1 summarizes the total number of each type of sample used for training and testing.

Table 1: Number of unique filter samples and liquid extract samples generated from extracts, dilutions of extracts and mixtures of extracts for each category of sample. We collected a total of 113 EEM spectra. Twelve of these spectra were used to generate training data leaving 101 spectra in the test set.

Sample Type	Number of Unique Filters	Samples for EEM	Spectra used for training
Cigarette	9	26	4
Diesel	10	29	4
Woodsmoke	9	26	4
Extract Mixtures	N/A*	21	0
Multiple-Exposure	5	5	0
Method Blanks	4	6	0
Total	**37**	**113**	**12**

*Mixtures of Cigarette, Diesel, and Woodsmoke samples

In addition to the samples collected in the laboratory, we collected twelve field samples to evaluate our method on real world samples. Eight field samples were taken in Seattle homes and in campus buildings ("background" field samples) and four were collected in areas we expected to be dominated by cigarette, diesel, or woodsmoke ("expected primary source" field samples). We collected and extracted the field samples using the same equipment and methods as the laboratory samples. EEM spectra collected from these two groups of field spectra are shown in Figure S7.

2.2 Fluorescence EEM analysis

PM extracts were stored in 4 ml vials (Cat # 66009-876 VWR, Edison, NJ) until analysis. For EEM spectroscopy ~3 ml of PM extract was transferred to a 1 cm x 1 cm quartz cuvette (Item # CV10Q3500FS, Thorlabs Inc., Newton, New Jersey). We collected EEM data using a fluorometer with an extended-UV 150W xenon-arc lamp (Aqualog-880-C, HORIBA Instruments Inc. Edison, New Jersey). We excited samples between 200 to 500 nm at 2 nm increments with an excitation slit width of 5 nm and recorded emission spectra between 246 and 826 nm on a CCD array. The CCD array has 1000 pixels each covering 0.58 nm. We collected data using 4-pixel binning giving an effective emission slit width of 2.32 nm. We kept emission data between 246 and 572 nm and excitation data between 224 and 500 nm. Emission data above 572 nm were discarded because minimal fluorescence was observed above this wavelength and excitation wavelengths below 224 nm were removed due to low excitation lamp intensity between 200 and 224 nm yielding extensive noise in the data. The raw fluorescent signal is corrected for detector response and lamp intensity by the instrument,[42] and is normalized to Raman units using Raman area data collected daily from a Milli-Q water sample.[43] Daily solvent blanks are recorded and used for blank subtraction to minimize the effect of Rayleigh and Raman scatter. To further reduce the effects of Rayleigh scatter we excised values within 10 nm of the first and second order Rayleigh scattering bands followed by replacement of the values using 2-dimmensional interpolation.[44] We did not correct for the inner filter effect because we observed absorbance below 0.2 for our samples that were recorded with the Aqualog during EEM collection.

2.3 Machine Learning for Identification of Sources Present

We used a CNN to identify the presence or absence of known PM sources in the EEM spectra. The CNN was trained on 6,375 training spectra generated from twelve single source spectra from cigarette, diesel, and woodsmoke PM (four from each source). We generated the training dataset using a data augmentation approach by mathematical combination of twelve original spectra, assuming fluorescence is linearly proportional to concentration (see section S3 for data supporting this assumption). The process for

generating spectra is shown schematically in Figure 1 and described in more detail in section S4. First, we created 1000 spectra for cigarette, diesel, and woodsmoke in a linearly spaced concentration range from 0 to 5 ug/mL resulting in 3000 single-source spectra. Then we created digital mixtures of the three sources in a logarithmically spaced concentration range from 0.01 to 6.3 in fifteen steps (15^3 combinations) giving 3375 training spectra consisting of mixtures. In creating mixtures by mathematically combining spectra from pure sources we assumed matrix effects of mixing to be negligible. We showed this to be a reasonable assumption by comparing digital and actual mixtures as illustrated in Figure S5.

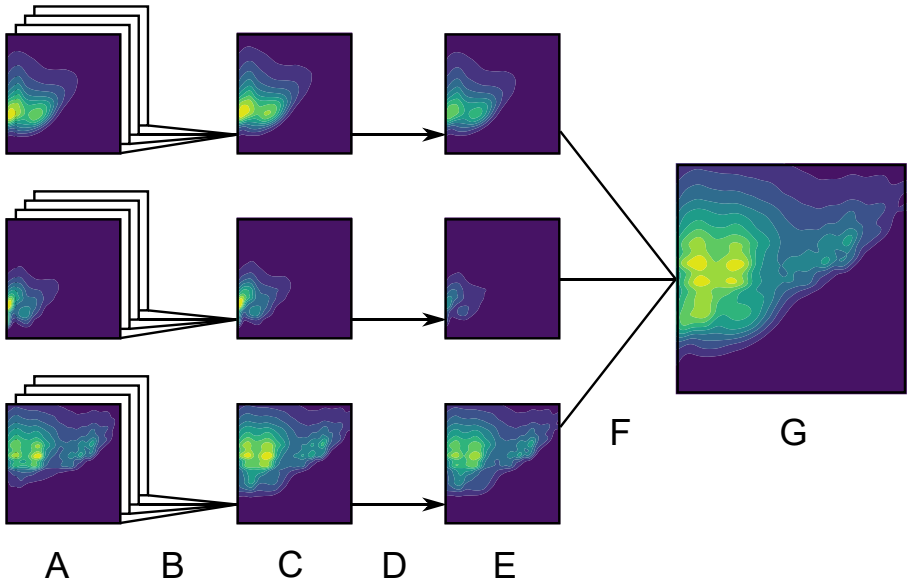


Figure 1: Graphical representation of the loop used to generate training spectra. Twelve original spectra, four from each of cigarette, diesel, and woodsmoke, (A) are averaged using random weights (B) producing prototypical spectra for each source (C) that are scaled (D) giving single source spectra at target concentration (E) that are combined (F) to generate a training spectra (G). This loop is repeated for the beginning for each training spectra generated in order to simulate variability associated with PM sampling and EEM collection.

The CNN used in this work consists of three convolutional layers each followed by max pooling,[45] as shown in Figure 2. All convolutions are performed using padding, so the dimensions of input and output

data are the same.[45] The first convolutional layer contains twenty 5-by-5 filters equating to 11.6 nm in emission (height) and 10 nm in excitation (width) as shown by the red box (Figure 2a). This is followed with 3-by-3 max pooling that reduces the data from 143-by-139 to 47-by-46. The second convolutional layer is ten 10-by-10 filters followed by 3-by-3 max pooling. The final convolutional layer applies ten 15-by-15 filters to the 15-by-15 feature maps. The output of the third convolutional layer is max pooled to a size of 5-by-5 and then flattened and connected to a dense neural network with three hidden layers having 512, 256, and 256 nodes in each layer, respectively. A dropout rate of 20% is used between all convolutional and fully connected layers.[46] The exponential linear unit was used as the activation function for all convolutional and hidden layers and a linear activation function was used for the output layer,[47] the loss function was the mean-squared-error.[29] The results described are from a network that was trained for 80 epochs. Details of how we selected the training duration are included in the SI. The CNN was implemented in Python 3 using Keras[48] and TensorFlow.[49]

Our EEMs are 2D spatial data made up of combinations of peaks and valleys, which correlate to a particular chemical or combination of chemicals that are extracted from the PM samples. These peaks and valleys vary in their intensity across emission and excitation dimensions. The convolution filters learn to fit these varying shapes to better detect peak presence as they are iteratively applied over the EEM. Subsequent convolutional layers are used to identify patterns of lower level features, for example, a second convolutional layer may look for a group of narrow peaks identified in the previous convolutional layer. The results of the convolutional layers are feature maps showing the presence or absence of features. The feature map data are fed into fully connected layers that map this information to the desired output. In our case this process assigns a predicted concentration value for each source that we use to predict the presence or absence of the sources.

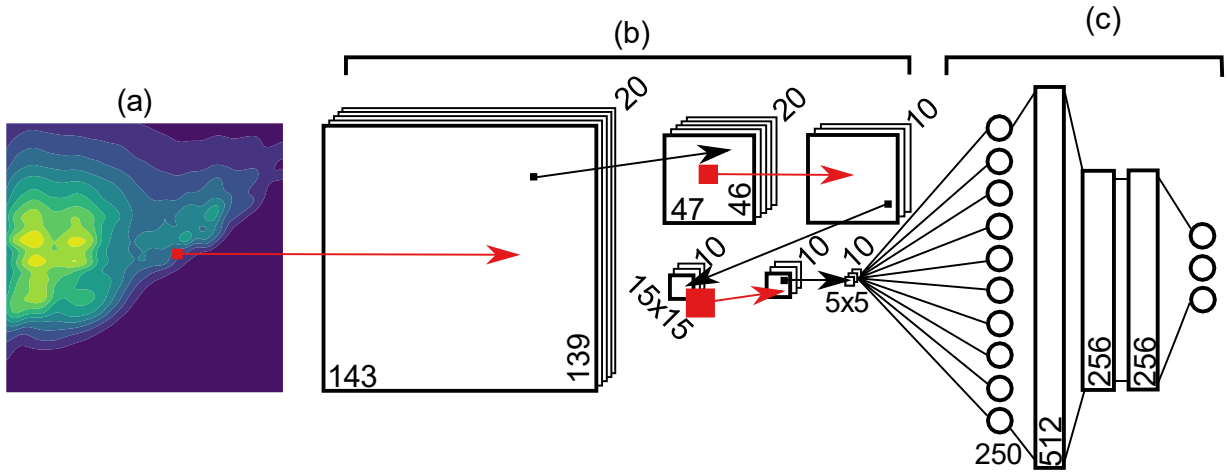


Figure 2: CNN Network Diagram. (a) Input spectra are first convolved with twenty 5-by-5 filters. (b) convolutional layers and max pooling layers are shown with associated data shapes. Convolutions (10x10 followed by 15x15) are shown in red and 3-by-3 max pooling is shown in black. Filters, data, and pooling sizes are shown to scale. (c) Output of the convolutional layers is flattened to a shape of 250 by 1 and fed into fully connected layers resulting in 3 output values (not to scale).

3. Results and Discussion

PM extracts from cigarette, woodsmoke, and diesel show unique EEM spectra, as shown in Figure 3. We also extracted cigarette and woodsmoke in methanol and water for comparison as shown in Figure S11. Cigarette spectra consist of two peaks at ~350 nm emission wavelength. Diesel spectra consist of single primary peak also located at ~350 nm emission with less fluorescence surrounding the peak than cigarette. Woodsmoke has spectra consisting of six peaks in the region from 400-475nm emission and 225 – 275 nm excitation. Cigarette and woodsmoke have similar maximum fluorescent intensity levels per mass of PM while diesel has a lower intensity. Woodsmoke shows fluorescence over the broadest region and generally at higher emission wavelengths than cigarette and diesel. The spectra from the different sources have overlapping regions suggesting challenges in distinguishing individual sources from mixed samples.

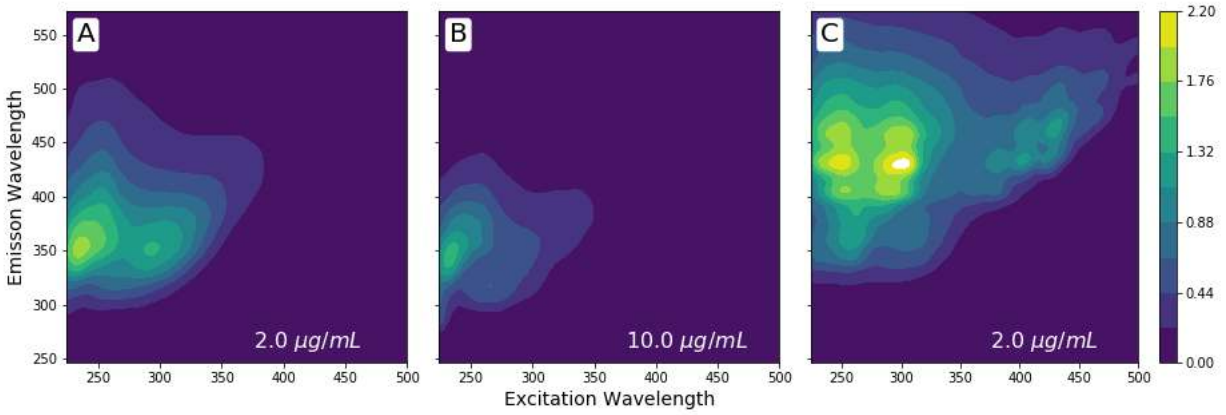


Figure 3: Fluorescence EEM spectra of (a) cigarette smoke at an extract concentration of 2 µg/ml, (b) diesel soot at 10 µg/ml, and (c) woodsmoke at 2 µg/ml. PM samples from the three sources were extracted in cyclohexane and exhibit unique spectral fingerprints. Cigarette and woodsmoke have similar maximum fluorescent intensity on a per mass basis while diesel has a lower signal intensity.

The CNN algorithm was trained on EEM spectra with known particulate concentrations generated as described in the methods section. True concentration values are determined from PM mass measurement and extraction volume. We then provide the CNN this same training data and ask it to predict the concentration of the three sources. These results are shown in Figure 4. The diagonal line represents perfect prediction of the samples where the CNN prediction values are equal to the values provided during training. The data points that result from the analysis of the original training data roughly follow the diagonal. The R^2 value for the fit to the training data for cigarette, diesel, and woodsmoke are 0.99, 0.97, 0.97 respectively. One reason for scatter in the training data is extracts at the same particulate matter concentration have different fluorescent signal strengths. This variation in signal strength is shown in Figure S4 which plots fluorescent intensity vs. concentration for single source spectra.

We then predict the concentration of the 101 test spectra which are shown in the parity plots of Figure 4. The results generally follow the diagonal trend, but there are significant under- and over-predictions. This can be attributed to the fact that total fluorescent intensity from a given source varies from sample to sample at the same concentration. The R^2 value for the fit to the test data for cigarette is

255 0.86, for diesel it is 0.79, and for woodsmoke it is 0.89. The lower R^2 values for the test data are due to the
256 overlap of the signals (Figure 4) that makes mixtures difficult to quantify and variation in fluorescent signal
257 intensity among samples at the same concentration (Figure S4).

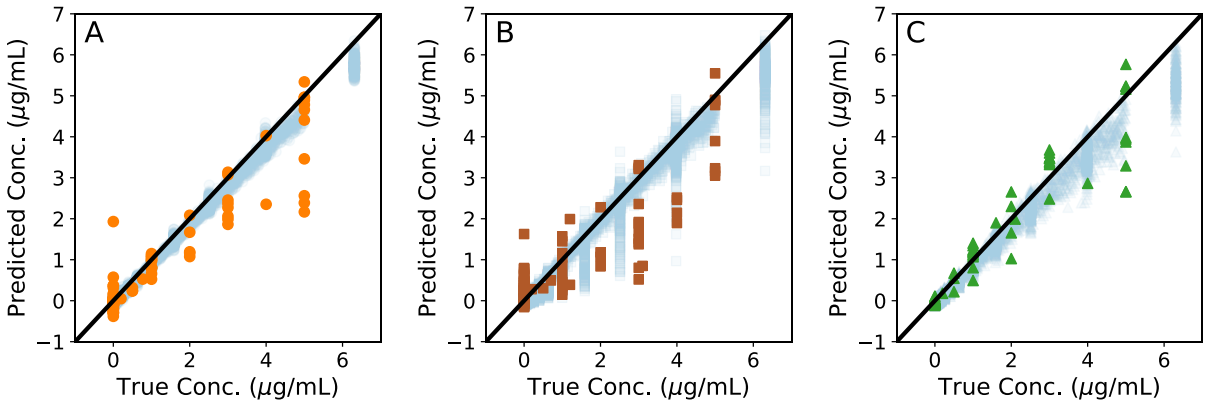


Figure 4: Parity plots showing predicted concentration vs. true extract concentrations for (a) cigarette (R^2 training = 0.99, R^2 test = 0.86), (b) diesel (R^2 training = 0.96, R^2 test = 0.79) and (c) woodsmoke (R^2 training = 0.97, R^2 test = 0.89). The data points shown as solid colors (orange, brown and green) are from 101 test spectra, the points shown in light blue are the results for the simulated training data.

258

259 We used the predicted results for samples containing only a single source to determine the limit of
260 detection (LoD). The LoD for each source was evaluated according to the Clinical and Laboratory
261 Standards Institute method as described in the supporting information.[50,51] The measured LoD for each
262 source is provided in Table 2. Here we provide the detection limit in mass of particulate matter per volume
263 of extraction liquid as well as a calculated particulate matter concentration per volume of sampled air,
264 assuming a 24-hour sampling time at 1.8 liters per minute (see section S1 for calculation details). Diesel
265 has the weakest fluorescence intensity and thus the highest LoD of 2.2 µg/mL cyclohexane or 2.6 µg/m^3
266 air. The LoD for each source in a 24-hour sampling period is significantly lower than the WHO and USEPA
267 24-hour mean exposure guidelines of 25 and 35 µg/m^3 respectively.[5,6]

268

269

Table 2: LoD determined by applying the CNN model to single source samples. The column reporting LoD in µg/mL is determined using PM mass measurement of filters dispersed in a volume of cyclohexane. The column reporting µg/m^3 in air is determined by converting the LoD in µg/mL to µg/m^3 assuming a 24-hours sampling time at an air sampling rate of 1.8 liters per minute.

Source	**LoD** [µg/mL cyclohexane]	**LoD** [µg/m^3 air]
Cigarette	0.6	0.7
Diesel	2.2	2.6
Woodsmoke	0.8	0.9

The ability to identify if PM from a source is present or absent above a threshold level could be a useful tool for clinicians and asthma patients in treating asthma or for asthma research, for example. To this end, we evaluated the ability of the CNN analysis of EEM spectra to detect the presence of individual sources above a threshold of 1 µg/mL. This threshold corresponds to an average exposure of nearly 10 µg/m^3, the WHO annual average guideline, during a three-hour sampling period at 1.8 L/min. In Figure 5, we plot the predicted concentration of each source in either a negative or positive column. Samples are considered positive if they had a true concentration (measured gravimetrically) of 1 µg/mL or greater of any of the single sources, and negative if they are below this concentration. This analysis method is based on the establishment of a cut off value for a qualitative diagnostic health test.[51] The clinical sensitivity, specificity, and overall accuracy of the diagnostic is then determined by choosing a threshold which delineates the positive from negative results. Depending on the purpose of the diagnostic test, the threshold may be set to achieve a specific outcome. For example, in the case of screening for a deadly but treatable disease, the number of false negatives would be minimized (i.e. maximizing sensitivity).[51] In this work, we choose the threshold that maximizes the accuracy for each source. Figure 5D shows a plot of source detection accuracy as a function of the calibrated threshold value used as a cut off between positive and negative detection. This plot shows that as we increase calibrated threshold value the detection accuracy for each source increases to a maximum and then decreases because as the threshold increases

nearly all positive samples are classified as negative. The threshold of maximum accuracy varies with the source. The predicted concentration thresholds for maximum detection accuracy for cigarette, diesel, and woodsmoke are 0.6, 0.8, and 0.7 µg/mL, respectively. These thresholds are shown by red horizontal lines in Figure 5A-C and when applied, we achieve an overall accuracy of 89%. The accuracies for identifying cigarette and woodsmoke were 98% and 99% respectively. Diesel was more challenging because of its low signal intensity relative to the other sources and had one false positive and seven false negatives giving an accuracy of 92%.

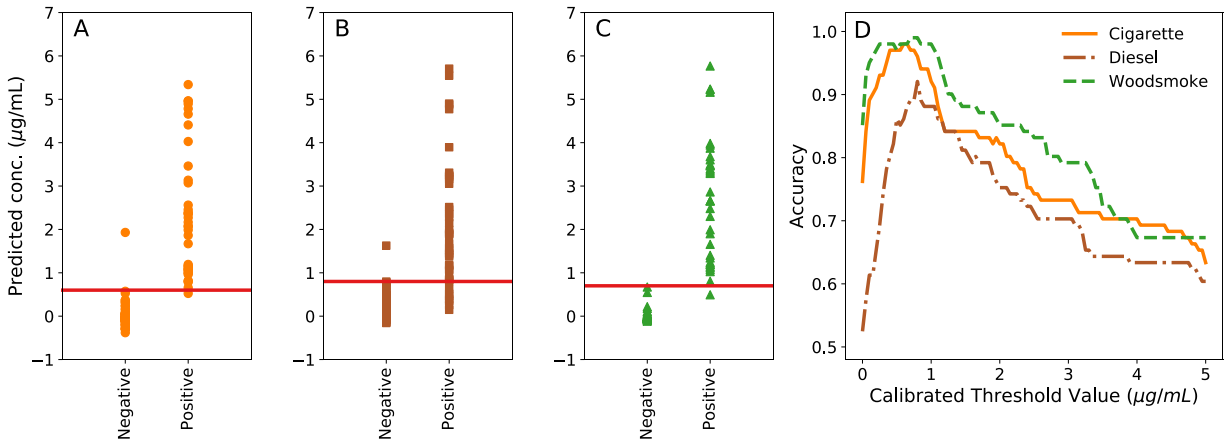


Figure 5: Classification plots showing classification of test data for (a) cigarette, (b) diesel and (c) woodsmoke sources as present or absent. Data points above the threshold (red horizontal line) are predicted as positive for the source. The location of the threshold was chosen to give the maximum accuracy for classifying each source individually. The source detection accuracies vs. calibrated thresholds are shown in (d).

After setting threshold values for source classification using all 101 test spectra, we evaluated the model performance on sub-groups of the test set. The sub-group of spectra containing single sources consisted of sixty-nine spectra from sixteen filter samples ranging in concentration from 0.2 µg/mL to 10 µg/mL. Within this group we classified the samples with an overall accuracy of 91%. Cigarette and woodsmoke spectra were identified with the best results while diesel was the most often misclassified with

303 a sensitivity of 0.91 and a specificity of 0.98. Next, we tested the algorithm on the sub-group of test spectra
304 containing two or more sources. Twenty-one samples were generated by mixing liquid extracts together
305 and five were from exposing an individual filter to multiple PM sources. The results of the analysis are
306 shown in Table 3. The CNN algorithm was able to identify the sources present in mixed samples with an
307 overall accuracy of 81%. The sensitivity and specificity for cigarette and woodsmoke was perfect; however,
308 diesel continued to show challenges with a specificity of unity and a sensitivity of 0.75. The relatively low
309 sensitivity of diesel is a result of the diesel spectra being weaker than and overlapping with the other sources
310 resulting in five false negative results. Finally, we evaluated six process blank spectra and the algorithm
311 correctly identified them all as not containing any of the sources. The results for classification of sub-groups
312 are summarized in Table 3.

Table 3: Classification results for sample sub-groups containing spectra with only one PM source and mixtures (two or three sources). The overall accuracy for the single source and mixtures groups were 91% and 81% respectively.

	Single Source			**Mixtures**			**Process Blanks**		
	Cigarette	Diesel	Wood	Cigarette	Diesel	Wood	Cigarette	Diesel	Wood
True +	16	21	17	21	16	20	0	0	0
True -	51	45	51	5	6	6	6	6	6
False +	1	1	0	0	0	0	0	0	0
False -	1	2	1	0	5	0	0	0	0
Accuracy	0.97	0.96	0.99	1.00	0.81	1.00	1.00	1.00	1.00
Sensitivity	0.94	0.91	0.94	1.00	0.75	1.00	1.00	1.00	1.00
Specificity	0.98	0.98	1.00	1.00	1.00	1.00	1.00	1.00	1.00

313

314 We evaluated a partial least squares (PLS) and a linear model to interpret the EEM spectra in the
315 same manner as the CNN. The linear model achieved an overall accuracy of 68% and the PLS model had
316 an overall accuracy of only 40%. Both the PLS and linear models performed poorly largely due to an
317 inability to accurately predict diesel concentration. We included the results of these comparative models
318 in section S7.

319

320 We tested the limits of our algorithm by applying it to a set of twelve field spectra. We had a

limited number of spectra available for this test and the algorithm was optimized for the laboratory samples, therefore the results are limited in scope to understanding potential difficulties when applying this method to a larger set of field samples. EEM spectra from the field samples were mathematically normalized to an extract concentration of 10 µg/mL for ease of interpreting results: the classification threshold used to train the CNN was 1 µg/mL so a spectrum will be classified as positive for a source if that source makes up 10% or more of the normalized EEM. We expect many sources of PM to contribute to the field samples such as crustal dust and biological material. The EEMs from the eight *background* field samples looked most similar to diesel and cigarette spectra (Figure S7A). The CNN predicted cigarette as present in three *background* samples, diesel as present in four and woodsmoke as present in two. We did not expect any of the *background* samples to contain cigarette smoke, as they came from non-smoking households and buildings, but this source was detected in three samples that had spectra of similar appearance to cigarette smoke. This suggests that some sources of PM have similar EEM spectra. Diesel may have been present as all samples were collected in urban areas of Seattle. Woodsmoke was detected in two *background* samples. The spectra where woodsmoke was detected looked most similar to diesel, but had higher fluorescent intensity than diesel at the same concentration (10 µg/mL). This illustrates that the CNN may give unexpected results when analyzing spectra that are different than spectra used in training. These results illustrate that an appropriate training set containing as many of the expected sources of PM as possible will be key for successful application to real world samples.

Woodsmoke was the expected primary source in an ambient sample taken in the UW cookstove lab and in a sample taken from an open window during a time when forest fire smoke was causing air pollution in Seattle. Woodsmoke was detected in the cookstove lab sample as expected and the EEM resembled other woodsmoke spectra (Figure S7B). We believe this was due to small amounts of woodsmoke escaping the ventilation system during stove testing. Woodsmoke was not detected in the sample taken during the forest fire smoke episode. We believe this is due to the forest fire smoke having a different composition than the laboratory generated woodsmoke due to a combination of aging during

atmospheric transport and different combustion conditions in a forest fire compared to a cookstove. Diesel exhaust was expected and detected as present in a sample taken in a mechanical room at the diesel exhaust exposure facility. We believe this was due to fugitive emissions of diesel exhaust as with the cookstove sample. Cigarette smoke was expected in a sample taken outdoors near a smoking area but only diesel was detected. In retrospect, we believe this was due to minimal amount of cigarette smoke present in the sample as the PM concentration measured by the filter was 6.3 $\mu g/m^3$, while the average concentration measured over the same time period at two nearby air monitoring sites in Seattle was 5.5 $\mu g/m^3$ showing that this outdoor sample likely consisted of a typical mixture of urban PM that would be expected to include diesel exhaust.[52]

4. Conclusions

We used a CNN model to successfully classify cigarette, diesel, and woodsmoke sources as present or absent in a series of laboratory samples. The limit of detection for our method is 0.7, 2.6, and 0.9 $\mu g/m^3$ in air for cigarette, diesel, and woodsmoke respectively. The CNN was able to identify cigarette and woodsmoke individually and in the presence of the other sources with 98% and 99% accuracy respectively, while classification of diesel was less accurate with an accuracy of 92%, sensitivity of 0.84 and specificity of 0.98. The overall classification accuracy for all three sources was 89%. When testing the limits of our algorithm by classifying field samples, some samples were classified as expected while in others, sources were detected as present even when they were not expected. This illustrates the need for a training data set with samples from more sources is needed for application to field samples.

Author Information:

Corresponding Author

*Phone: +1 (206) 543-9834; fax: +1 (206) 685-8047; e-mail: jposner@uw.edu

Notes:

The authors declare no competing financial interest.

Acknowledgements:

373 This work was funded by the National Institute of Biomedical Imaging and Bioengineering (NIBIB) grant
374 U01 EB021923. Data analysis was supported by the Data Intensive Research Enabling Clean Technology
375 (DIRECT) NSF National Research Traineeship (DGE-1633216). We thank Ben Sullivan, Garrett Allawatt,
376 and Devin Udesen of the UW Clean Cookstoves Lab for their assistance collecting woodsmoke samples,
377 Jim Stewart for his help collecting diesel samples, and Tim Gould for his guidance on weighing filters and
378 sample collection.
379
380

References:

(1) Stanaway, J. D.; Afshin, A.; Gakidou, E.; Lim, S. S.; Abate, D.; Abate, K. H.; Abbafati, C.; Abbasi, N.; Abbastabar, H.; Abd-Allah, F.; et al. Global, Regional, and National Comparative Risk Assessment of 84 Behavioural, Environmental and Occupational, and Metabolic Risks or Clusters of Risks for 195 Countries and Territories, 1990–2017: A Systematic Analysis for the Global Burden of Disease Study 2017. *The Lancet* **2018**, *392* (10159), 1923–1994. https://doi.org/10.1016/S0140-6736(18)32225-6.

(2) Godish, T. *Air Quality*, 3rd ed.; CRC/Lewis Publishers: Boca Raton, Fla., 1997.

(3) Tsuda, A.; Henry, F. S.; Butler, J. P. Particle Transport and Deposition: Basic Physics of Particle Kinetics. *Compr. Physiol.* **2013**, *3* (4), 1437–1471. https://doi.org/10.1002/cphy.c100085.

(4) Miller, F. J.; Gardner, D. E.; Graham, J. A.; Lee, R. E.; Wilson, W. E.; Bachmann, J. D. Size Considerations for Establishing a Standard for Inhalable Particles. *J. Air Pollut. Control Assoc.* **1979**, *29* (6), 610–615. https://doi.org/10.1080/00022470.1979.10470831.

(5) US EPA, O. Table of Historical Particulate Matter (PM) National Ambient Air Quality Standards (NAAQS) https://www.epa.gov/pm-pollution/table-historical-particulate-matter-pm-national-ambient-air-quality-standards-naaqs (accessed May 3, 2018).

(6) World Health Organization. Regional Office for Europe. *Air Quality Guidelines: Global Update 2005 : Particulate Matter, Ozone, Nitrogen Dioxide, and Sulfur Dioxide.*; World Health Organization Europe: Copenhagen, Denmark, 2006.

(7) Janssen, N. A. H.; Hoek, G.; Simic-Lawson, M.; Fischer, P.; Van Bree, L.; Ten Brink, H.; Keuken, M.; Atkinson, R. W.; Anderson, H. R.; Brunekreef, B.; et al. Black Carbon as an Additional Indicator of the Adverse Health Effects of Airborne Particles Compared with PM10 and PM2.5. *Environ. Health Perspect.* **2011**, *119* (12), 1691–1699. https://doi.org/10.1289/ehp.1003369.

(8) Bell, M. L.; Ebisu, K.; Leaderer, B. P.; Gent, J. F.; Lee, H. J.; Koutrakis, P.; Wang, Y.; Dominici, F.; Peng, R. D. Associations of $PM_{2.5}$ Constituents and Sources with Hospital Admissions: Analysis of Four Counties in Connecticut and Massachusetts (USA) for Persons ≥ 65 Years of Age. *Environ. Health Perspect.* **2014**, *122* (2), 138–144. https://doi.org/10.1289/ehp.1306656.

(9) Stanek, L. W.; Sacks, J. D.; Dutton, S. J.; Dubois, J.-J. B. Attributing Health Effects to Apportioned Components and Sources of Particulate Matter: An Evaluation of Collective Results. *Atmos. Environ.* **2011**, *45* (32), 5655–5663. https://doi.org/10.1016/j.atmosenv.2011.07.023.

(10) West, J. J.; Cohen, A.; Dentener, F.; Brunekreef, B.; Zhu, T.; Armstrong, B.; Bell, M. L.; Brauer, M.; Carmichael, G.; Costa, D. L.; et al. "What We Breathe Impacts Our Health: Improving Understanding of the Link between Air Pollution and Health." *Environ. Sci. Technol.* **2016**, *50* (10), 4895–4904. https://doi.org/10.1021/acs.est.5b03827.

(11) Adams, K.; Greenbaum, D. S.; Shaikh, R.; Erp, A. M. van; Russell, A. G. Particulate Matter Components, Sources, and Health: Systematic Approaches to Testing Effects. *J. Air Waste Manag. Assoc.* **2015**, *65* (5), 544–558. https://doi.org/10.1080/10962247.2014.1001884.

(12) Hime, N.; Marks, G.; Cowie, C.; Hime, N. J.; Marks, G. B.; Cowie, C. T. A Comparison of the Health Effects of Ambient Particulate Matter Air Pollution from Five Emission Sources. *Int. J. Environ. Res. Public. Health* **2018**, *15* (6), 1206. https://doi.org/10.3390/ijerph15061206.

(13) Koenig, J. Q. Air Pollution and Asthma. *J. Allergy Clin. Immunol.* **1999**, *104* (4), 717–722. https://doi.org/10.1016/S0091-6749(99)70280-0.

(14) National Asthma Education and Prevention Program. Expert Panel Report 3 (EPR-3): Guidelines for the Diagnosis and Management of Asthma–Summary Report 2007. *J. Allergy Clin. Immunol.* **2007**, *120* (5), S94–S138. https://doi.org/10.1016/j.jaci.2007.09.029.
(15) Lewis, S. A.; Antoniak, M.; Venn, A. J.; Davies, L.; Goodwin, A.; Salfield, N.; Britton, J.; Fogarty, A. W. Secondhand Smoke, Dietary Fruit Intake, Road Traffic Exposures, and the Prevalence of Asthma: A Cross-Sectional Study in Young Children. *Am. J. Epidemiol.* **2005**, *161* (5), 406–411. https://doi.org/10.1093/aje/kwi059.
(16) Duncan, G. E.; Seto, E.; Avery, A. R.; Oie, M.; Carvlin, G.; Austin, E.; Shirai, J. H.; He, J.; Ockerman, B.; Novosselov, I. Usability of a Personal Air Pollution Monitor: Design-Feedback Iterative Cycle Study. *JMIR MHealth UHealth* **2018**, *6* (12), e12023. https://doi.org/10.2196/12023.
(17) Moerner, W. E.; Fromm, D. P. Methods of Single-Molecule Fluorescence Spectroscopy and Microscopy. *Rev. Sci. Instrum.* **2003**, *74* (8), 3597–3619. https://doi.org/10.1063/1.1589587.
(18) Elcoroaristizabal, S.; de Juan, A.; García, J. A.; Durana, N.; Alonso, L. Comparison of Second-Order Multivariate Methods for Screening and Determination of PAHs by Total Fluorescence Spectroscopy. *Chemom. Intell. Lab. Syst.* **2014**, *132*, 63–74. https://doi.org/10.1016/j.chemolab.2014.01.005.
(19) Nahorniak, M. L.; Booksh, K. S. Excitation-Emission Matrix Fluorescence Spectroscopy in Conjunction with Multiway Analysis for PAH Detection in Complex Matrices. *Analyst* **2006**, *131* (12), 1308–1315. https://doi.org/10.1039/B609875D.
(20) Johnson, D. W.; Callis, J. B.; Christian, G. D. Rapid Scanning Fluorescence Spectroscopy. *Anal. Chem.* **1977**, *49* (8), 747A-757A. https://doi.org/10.1021/ac50016a769.
(21) Andrade-Eiroa, Á.; Canle, M.; Cerdá, V. Environmental Applications of Excitation-Emission Spectrofluorimetry: An In-Depth Review II. *Appl. Spectrosc. Rev.* **2013**, *48* (2), 77–141. https://doi.org/10.1080/05704928.2012.692105.
(22) Aryal, R.; Lee, B.-K.; Beecham, S.; Kandasamy, J.; Aryal, N.; Parajuli, K. Characterisation of Road Dust Organic Matter as a Function of Particle Size: A PARAFAC Approach. *Water. Air. Soil Pollut.* **2015**, *226* (2), 24. https://doi.org/10.1007/s11270-014-2289-y.
(23) Chen, Q.; Miyazaki, Y.; Kawamura, K.; Matsumoto, K.; Coburn, S.; Volkamer, R.; Iwamoto, Y.; Kagami, S.; Deng, Y.; Ogawa, S.; et al. Characterization of Chromophoric Water-Soluble Organic Matter in Urban, Forest, and Marine Aerosols by HR-ToF-AMS Analysis and Excitation–Emission Matrix Spectroscopy. *Environ. Sci. Technol.* **2016**, *50* (19), 10351–10360. https://doi.org/10.1021/acs.est.6b01643.
(24) Elcoroaristizabal, S.; Juan, A. de; García, J. A.; Elorduy, I.; Durana, N.; Alonso, L. Chemometric Determination of PAHs in Aerosol Samples by Fluorescence Spectroscopy and Second-Order Data Analysis Algorithms. *J. Chemom.* **2014**, *28* (4), 260–271. https://doi.org/10.1002/cem.2604.
(25) Matos, J. T. V.; Freire, S. M. S. C.; Duarte, R. M. B. O.; Duarte, A. C. Natural Organic Matter in Urban Aerosols: Comparison between Water and Alkaline Soluble Components Using Excitation–Emission Matrix Fluorescence Spectroscopy and Multiway Data Analysis. *Atmos. Environ.* **2015**, *102*, 1–10. https://doi.org/10.1016/j.atmosenv.2014.11.042.
(26) Mladenov, N.; Alados-Arboledas, L.; Olmo, F. J.; Lyamani, H.; Delgado, A.; Molina, A.; Reche, I. Applications of Optical Spectroscopy and Stable Isotope Analyses to Organic Aerosol Source Discrimination in an Urban Area. *Atmos. Environ.* **2011**, *45* (11), 1960–1969. https://doi.org/10.1016/j.atmosenv.2011.01.029.
(27) Mladenov, N.; Lípez-Ramos, J.; McKnight, D. M.; Rechea, I. Alpine Lake Optical

Properties as Sentinels of Dust Deposition and Global Change. *Limnol. Oceanogr.* **2009**, *54* (6part2), 2386–2400. https://doi.org/10.4319/lo.2009.54.6_part_2.2386.
(28) Nakajima, H.; Okada, K.; Kuroki, Y.; Nakama, Y.; Handa, D.; Arakaki, T.; Tanahara, A. Photochemical Formation of Peroxides and Fluorescence Characteristics of the Water-Soluble Fraction of Bulk Aerosols Collected in Okinawa, Japan. *Atmos. Environ.* **2008**, *42* (13), 3046–3058. https://doi.org/10.1016/j.atmosenv.2007.12.045.
(29) Hastie, T. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Second edition, [corrected at 7th printing].; Springer series in statistics; Springer: New York, NY, USA, 2013.
(30) Goodfellow, I.; Bengio, Y.; Courville, A. *Deep Learning*; MIT Press, 2016.
(31) Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *ArXiv14091556 Cs* **2014**.
(32) Krizhevsky, A.; Sutskever, I.; Hinton, G. E. ImageNet Classification with Deep Convolutional Neural Networks. In *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1*; NIPS'12; Curran Associates Inc.: USA, 2012; pp 1097–1105.
(33) Di, Q.; Kloog, I.; Koutrakis, P.; Lyapustin, A.; Wang, Y.; Schwartz, J. Assessing PM2.5 Exposures with High Spatiotemporal Resolution across the Continental United States. *Environ. Sci. Technol.* **2016**, *50* (9), 4712–4721. https://doi.org/10.1021/acs.est.5b06121.
(34) Song, X.-H.; Hopke, P. K. Solving the Chemical Mass Balance Problem Using an Artificial Neural Network. *Environ. Sci. Technol.* **1996**, *30* (2), 531–535. https://doi.org/10.1021/es950281o.
(35) Carstea, E. M.; Baker, A.; Bieroza, M.; Reynolds, D. Continuous Fluorescence Excitation–Emission Matrix Monitoring of River Organic Matter. *Water Res.* **2010**, *44* (18), 5356–5366. https://doi.org/10.1016/j.watres.2010.06.036.
(36) Bieroza, M.; Baker, A.; Bridgeman, J. Exploratory Analysis of Excitation-Emission Matrix Fluorescence Spectra with Self-Organizing Maps as a Basis for Determination of Organic Matter Removal Efficiency at Water Treatment Works. *J. Geophys. Res. Biogeosciences* **2009**, *114* (G4). https://doi.org/10.1029/2009JG000940.
(37) García-Reiriz, A.; Damiani, P. C.; Olivieri, A. C.; Cañada-Cañada, F.; Muñoz de la Peña, A. Nonlinear Four-Way Kinetic-Excitation−Emission Fluorescence Data Processed by a Variant of Parallel Factor Analysis and by a Neural Network Model Achieving the Second-Order Advantage: Malonaldehyde Determination in Olive Oil Samples. *Anal. Chem.* **2008**, *80* (19), 7248–7256. https://doi.org/10.1021/ac8007829.
(38) García-Reiriz, A.; Damiani, P. C.; Olivieri, A. C. Analysis of Amoxicillin in Human Urine by Photo-Activated Generation of Fluorescence Excitation–Emission Matrices and Artificial Neural Networks Combined with Residual Bilinearization. *Anal. Chim. Acta* **2007**, *588* (2), 192–199. https://doi.org/10.1016/j.aca.2007.02.020.
(39) Sullivan, B.; Allawatt, G.; Emery, A.; Means, P.; Kramlich, J.; Posner, J. Time-Resolved Particulate Emissions Monitoring of Cookstove Biomass Combustion Using a Tapered Element Oscillating Microbalance. *Combust. Sci. Technol.* **2017**, *189* (6), 923–936. https://doi.org/10.1080/00102202.2016.1253564.
(40) Gould, T.; Larson, T.; Stewart, J.; Kaufman, J. D.; Slater, D.; Mcewen, N. A Controlled Inhalation Diesel Exhaust Exposure Facility with Dynamic Feedback Control of PM Concentration. *Inhal. Toxicol.* **2008**, *20* (1), 49–52. https://doi.org/10.1080/08958370701758478.
(41) Allen, R.; Box, M.; Liu, L.-J. S.; Larson, T. V. A Cost-Effective Weighing Chamber for

519 Particulate Matter Filters. *J. Air Waste Manag. Assoc.* **2001**, *51* (12), 1650–1653.
520 https://doi.org/10.1080/10473289.2001.10464392.
521 (42) Engelborghs, Y.; Visser, A. J. W. G. *Fluorescence Spectroscopy and Microscopy:*
522 *Methods and Protocols*; Methods in molecular biology (Clifton, N.J.) ; v. 1076; Humana Press:
523 New York, 2014.
524 (43) Murphy, K. R.; Butler, K. D.; Spencer, R. G. M.; Stedmon, C. A.; Boehme, J. R.; Aiken,
525 G. R. Measurement of Dissolved Organic Matter Fluorescence in Aquatic Environments: An
526 Interlaboratory Comparison. *Environ. Sci. Technol.* **2010**, *44* (24), 9405–9412.
527 https://doi.org/10.1021/es102362t.
528 (44) Zepp, R. G.; Sheldon, W. M.; Moran, M. A. Dissolved Organic Fluorophores in
529 Southeastern US Coastal Waters: Correction Method for Eliminating Rayleigh and Raman
530 Scattering Peaks in Excitation–Emission Matrices. *Mar. Chem.* **2004**, *89* (1), 15–36.
531 https://doi.org/10.1016/j.marchem.2004.02.006.
532 (45) Dumoulin, V.; Visin, F. A Guide to Convolution Arithmetic for Deep Learning.
533 *ArXiv160307285 Cs Stat* **2016**.
534 (46) Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; Salakhutdinov, R. Dropout: A
535 Simple Way to Prevent Neural Networks from Overfitting. *J Mach Learn Res* **2014**, *15* (1),
536 1929–1958.
537 (47) Clevert, D.-A.; Unterthiner, T.; Hochreiter, S. Fast and Accurate Deep Network Learning
538 by Exponential Linear Units (ELUs). *ArXiv151107289 Cs* **2015**.
539 (48) Chollet, F. Keras https://keras.io (accessed Sep 21, 2018).
540 (49) Abadi, M.; Agarwal, A.; Barham, P.; Brevdo, E.; Chen, Z.; Citro, C.; Corrado, G. S.;
541 Davis, A.; Dean, J.; Devin, M.; et al. *TensorFlow: Large-Scale Machine Learning on*
542 *Heterogeneous Systems*; 2015.
543 (50) Tholen, D. W. *Protocols for Determination of Limits of Detection and Limits of*
544 *Quantitation: Approved Guideline*; NCCLS: Wayne, Pa., 2004.
545 (51) Borysiak, M. D.; Thompson, M. J.; Posner, J. D. Translating Diagnostic Assays from the
546 Laboratory to the Clinic: Analytical and Clinical Metrics for Device Development and
547 Evaluation. *Lab. Chip* **2016**, *16* (8), 1293–1313. https://doi.org/10.1039/c6lc00015k.
548 (52) Puget Sound Clean Air Agency - Air Graphing Tool
549 https://secure.pscleanair.org/airgraphing (accessed Jan 22, 2019).
550

Supporting Information

Excitation Emission Matrix Fluorescence Spectroscopy for Combustion Generated Particulate Matter Source Identification

Jay W. Rutherford,[1] Neal Dawson-Elli,[1] Anne. M. Manicone,[5] Gregory V. Korshin,[2] Igor V. Novosselov,[2] Edmund Seto,[4] Jonathan D. Posner[1,2,3]

[1]Department of Chemical Engineering, [2]Department of Mechanical Engineering, [3]Department of Family Medicine, [4]Environmental and Occupational Health Sciences, [5]Department of Medicine: Pulmonary, Critical Care and Sleep Medicine University of Washington.

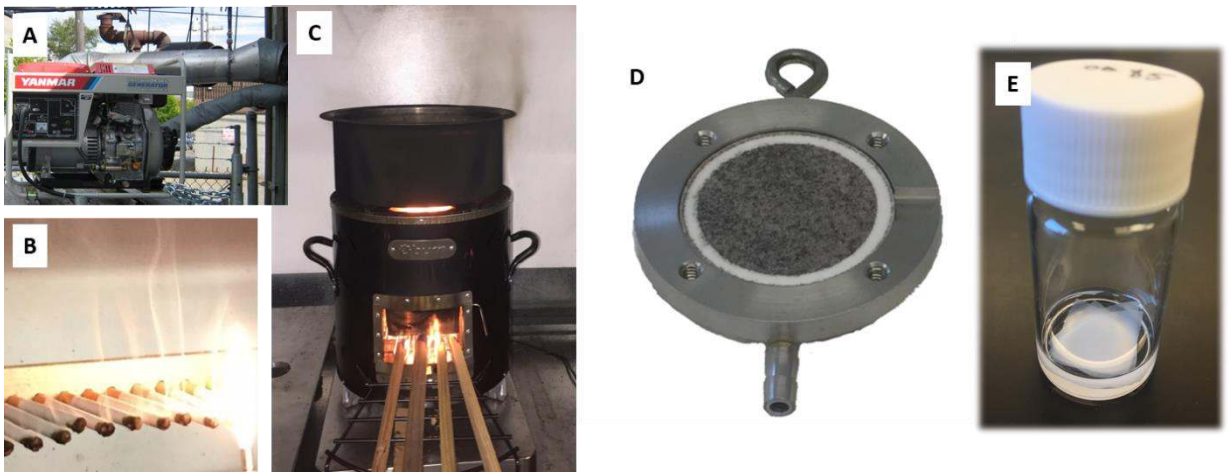


Figure S1. Laboratory sources of particulate matter: (a) diesel exposure chamber, (b) cigarettes, and (c) clean cookstove. Particulate matter collected on a 1-inch diameter PTFE filter shown in a HPEM with the inlet removed(d) and PTFE filter in cyclohexane for extraction (e).

S1 Limit of Detection

We determined the limit of detection (LoD) of pure sources using the Clinical and Laboratory Standards Institute method for determining LoD. [1,2] This procedure accounts for variation present in blank measurements and measurements of low levels of analyte to assign a LoD that represents 95% confidence in differentiating a low concentration sample from a blank and visa-versa. First the limit of blank (LoB) was determined as,

$$LoB = \mu_B + 1.645\sigma_B, \qquad \textit{(Equation S1)}$$

where μ_B is the mean and σ_B is the standard deviation of the blank measurement. This method is assuming measurements are normally distributed. 1.645 is the z-score for which 95% of values in the standard normal distribution are below. Next, dilutions of extracts at low levels were analyzed. Extract concentrations were 0.5, 1.0, and 2.0 µg/mL for cigarette and woodsmoke and 1.0, 2.0 and 3.0 for diesel. From these low-level measurements, the LoD was calculated as,

$$LoD = LoB + 1.645/\left(1 - \frac{1}{4f}\right) \cdot \sigma_S, \qquad \textit{(Equation S2)}$$

where σ_S is the average standard deviation of low-level sample measurements and f is the degrees of freedom calculated as the number of low-level samples analyzed minus one. For cigarette and woodsmoke, σ_S was calculated from the 0.5, 1.0 and 2.0 µg/mL samples and for diesel from the 1.0, 2.0 and 3.0 µg/mL samples. Figure S2 shows error bars calculated as $1.645/\left(1 - \frac{1}{4f}\right) \cdot \sigma_S$ for all the low-level samples measured. Graphically, the LoD is where these error bars intersect with the LoB which is shown as a black horizontal dashed line in Figure S2. The LoD in terms of actual concentration (shown on the x-axis) is determined by finding the value at which the line of best fit (shown with a dotted line in Figure S2) intersects the value of LoD. The LoD as reported in Table 2 is shown in Figure S2 by the red dashed vertical line.

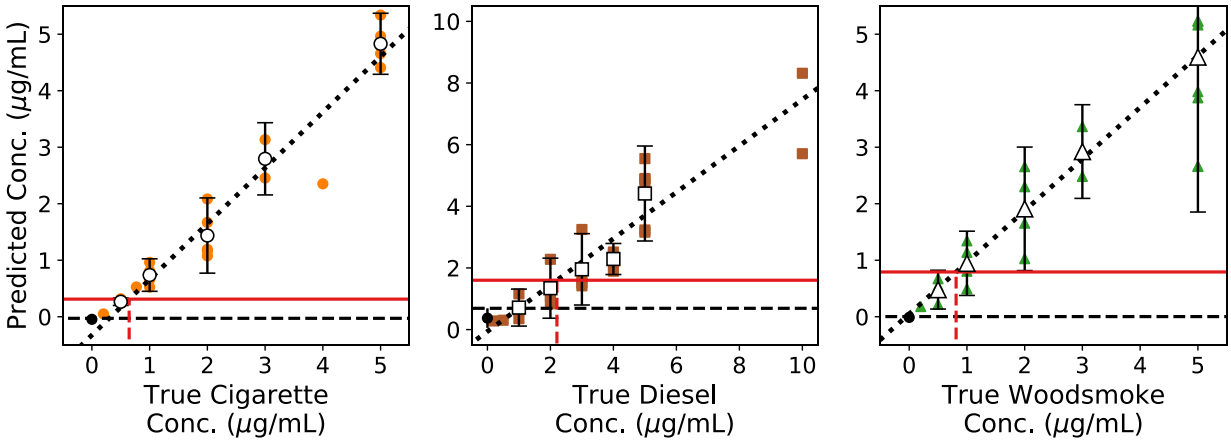


Figure S2. Plots showing data used to determine LoD. Only samples containing a single source were used to determine the LoD. The LoB (Equation S1) is shown by the black horizontal dashed line, the LoD (Equation S2) on the predicted concentration-axis (y-axis) is shown by the red horizontal line. The LoD on the true concentration axis is shown by the vertical red dashed line. This value (reported in Table 2) is determined using the line of best fit which is shown as a dotted line. The model is not constrained to predicted only positive concentrations so the LoB for cigarette is slightly below zero due to predicted concentrations of blank samples being just below zero.

The Lodes in units of volume of air sampled were calculated assuming a sampling time of 24 hours, an air sampling rate of 1.8 liters/min and an extraction volume of 3 ml. An example conversion is given as,

$$1\frac{\mu g}{ml\ solvent} * \frac{3\ ml\ solvent}{sample} * \frac{sample}{24\ hours} * \frac{1\ hour}{60\ mins} * \frac{1\ min}{1.8\ liters\ air} * \frac{1000\ liters}{1\ m^3} = 1.16\frac{\mu g}{m^3}.$$

S2 Accuracy, Sensitivity and Specificity

The individual classification accuracy and overall classification accuracy are calculated as: Accuracy = (Correctly Identified Samples) / (Total Samples). Individual accuracy only considers the classification results for the source of concern while overall accuracy considers all sources simultaneouslty such that if any of the three sources are incorrectly identified (false positive or false negative) the entire sample is considered incorrect. Sensitivity is a measure of the ability of a test to identify positive results successfully defined as: Sensitivity = (True Positives) / (True Positives + False Negatives). Specificity is a measure of ability of a test to classify negative results properly as defined by: Specificity = (True Negatives) / (True negatives + False Positives).

S3 Fluorescent Intensity

The measured fluorescent intensity is linearly proportional to PM extract concentration for an individual filter. Figure S3 plots integrated fluorescence intensity for single source spectra vs. PM extract concentration. The R^2 values for these dilution series are all close to one. This nearly perfect linear relationship for a dilution series supports our approach of linearly scaling fluorescent intensity when generating training data.

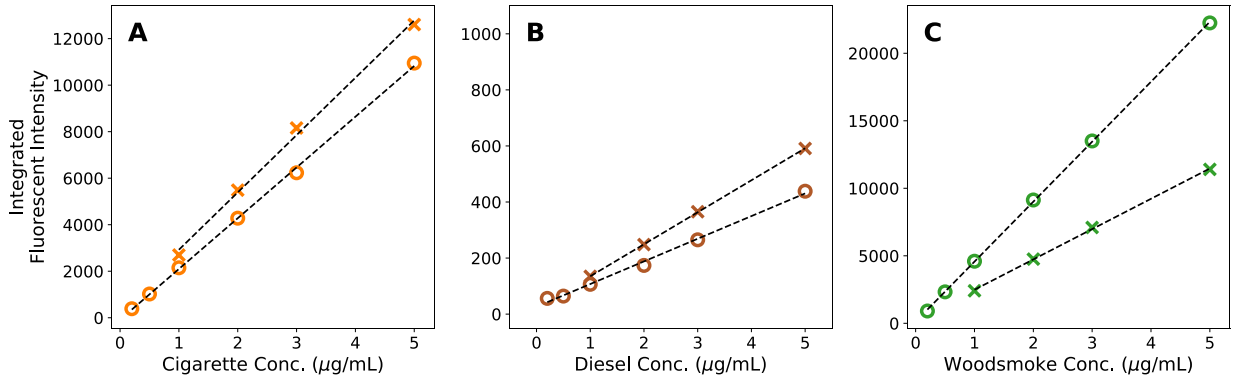


Figure S3: Integrated fluorescence intensity for single source spectra vs. PM extract concentration. Two dilution series are shown for each source (marked with circles and X's) for cigarette (a) diesel (b) and woodsmoke (c). The R^2 values for linear fit to these dilution series are all above 0.998. These high R^2 values support our method of linearly scaling fluorescent intensity when generating training data.

Although fluorescence intensity scales linearly with concentration for a single filter, there is significant variation in fluorescence from filter to filter for a given mass concentration. The fluorescence coefficient of variation for samples at 5 µg/mL is 0.09, 0.27, and 0.66 for the cigarette, diesel, and woodsmoke. For this reason, if we fit a single linear line to all single source spectra for each source, we obtain R^2 values of 0.97, 0.84, and 0.69 for cigarette, diesel, and woodsmoke, respectively (Figure S4). We believe this variation is due in part to the variability in source concentration in the air sampled as well as variability in combustion conditions. For example, we have observed that the fluorescent intensity from a filter decreases when HEPA filtered air is passed through the filter after sampling the source.

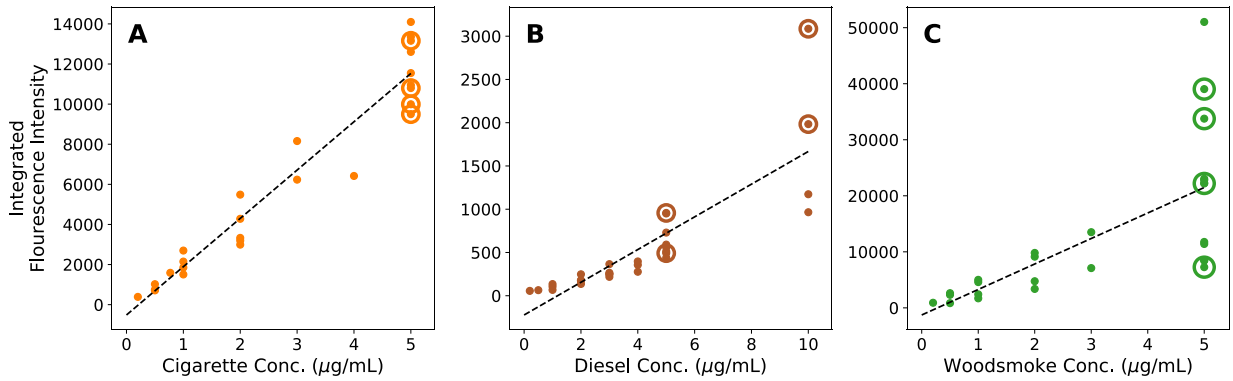


Figure S4: Soot mass vs. integrated fluorescence intensity. Cigarette (A) and woodsmoke (C) have higher fluorescence per mass than diesel soot (B). Samples from multiple filters show a positive correlation with concentration, r^2=0.97 for cigarette (26 spectra), 0.84 for diesel (29 spectra) and 0.69 for wood smoke (26 spectra). The four points in each plot with a circle around them are the samples used for generating the training data.

S4 Data Augmentation

We generated an augmented training dataset consisting of 6,375 spectra by linearly scaling and mathematically combining spectra from 12 unique samples (4 from each of the 3 sources, cigarette, diesel and woodsmoke). These 12 spectra are only used in the training data and are not used for testing. Figure S4 shows the integrated fluorescent intensity of single source spectra for cigarette, diesel and woodsmoke. The spectra chosen for the data augmentation process are denoted with large circles in Figure S4. These spectra were chosen so that the samples used to generate the training data are not present in any samples in the test data. This means the samples used for data augmentation are not present in any dilutions or mixtures. For cigarette these spectra lay roughly on the line of best fit (Figure S4-A), for diesel three of the calibration spectra lie roughly on the line of best fit and one has above average signal intensity (Figure S4-B) and for woodsmoke two calibration spectra are above the line of best fit, one below and one near (Figure S4-C). The relative intensity of the spectra chosen for training influences the model to over or underpredict the test spectra. For example, the woodsmoke calibration spectra above average fluorescent intensity in total so the model tends to underpredict the concentrations of woodsmoke in the test data.

To simulate noise and variability we average the four spectra from each source together using a weighted average using randomly assigned weights to produce a prototypical spectrum for each source. The prototypical spectra for the three sources are scaled to the desired concentration and combined to generate a training spectrum. For each spectrum generated the entire process is repeated in a loop to simulate variability associated with PM sampling and EEM collection. This process is depicted graphically in Figure 1.

We employed two sampling strategies to create the augmented training dataset. First, we created 1000 spectra for each of the three sources containing only one source in a linearly spaced concentration range from 0 to 5 ug/mL, resulting in 3000 spectra. Then we created digital mixtures of the three sources in a logarithmically spaced concentration range from 0.01 to 6.3 in fifteen steps (15^3 combinations) giving 3375 training spectra consisting of mixtures.

This process assumes the Beer–Lambert law to apply to absorbance, constant fluorescence quantum yield, and negligible inner filter and matrix effects.[3] We tested these assumptions in 2 ways. First, we analyzed series of dilutions made using extracts from single filters. When looking at just these dilution series as plotted in Figure S3 we show that fluorescence is indeed proportional to concentration as we are assuming. Next, to test that matrix effects due to mixing multiple sources together are minimal we compared liquid and digital mixtures of the 3 sources to confirm that liquid mixture and the digital mixture gave similar results. The results of this analysis are shown in Figure S5.

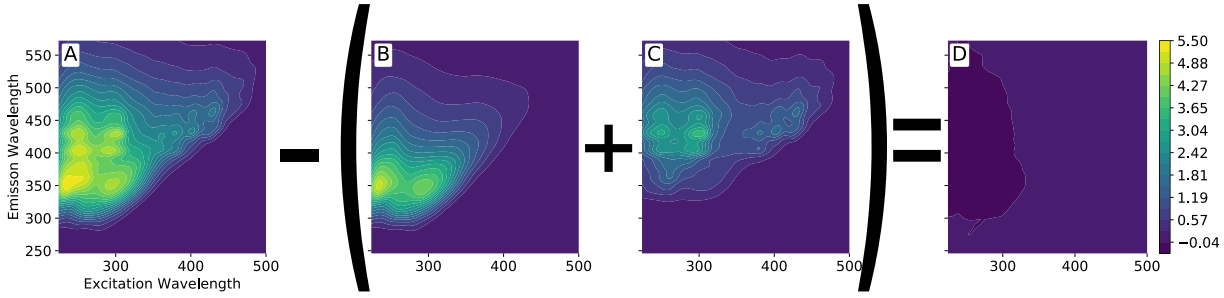
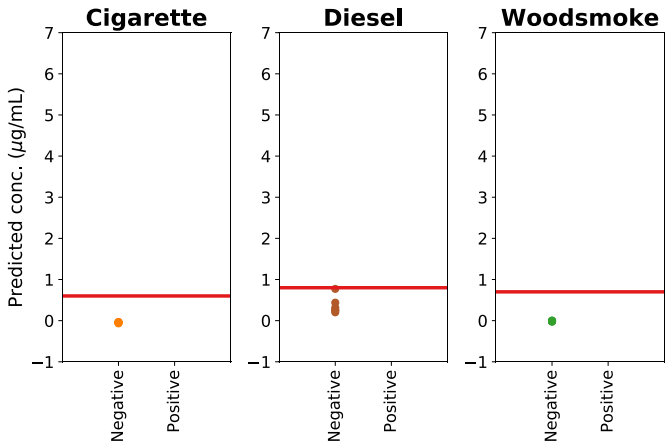


Figure S5: (a) shows a liquid mixture of cigarette and woodsmoke extracts each at 5 µg/mL, (b) and (c) show cigarette and woodsmokes extracts respectively also at 5 µg/mL, and (d) is the result of subtracting (b) and (c) from (a) showing nearly zero remaining signal which illustrates that matrix effects due to mixing are minimal.

S5 Classification of Method Blanks

We collected spectra from four unique method blank filters. Extracts from two of these method blanks were scanned on two separate days giving a total of 6 method blank spectra. The method blank filters were processed in the same manner as other filters which included weighing, loading intro sampling devices, and extraction as described in the materials and methods section. The only step that was not conducted with these samples was air sampling. The spectra from these samples showed almost no fluorescence and when input to the model the classification results (Figure S6) were all negative showing our sampling procedure and filter material does not cause significant interference.



	Cigarette	Diesel	Woodsmoke
True+	0	0	0
True-	6	6	6
False+	0	0	0
False-	0	0	0
Accuracy	1.00	1.00	1.00
Sensitivity	nan	nan	nan
Specificity	1.00	1.00	1.00

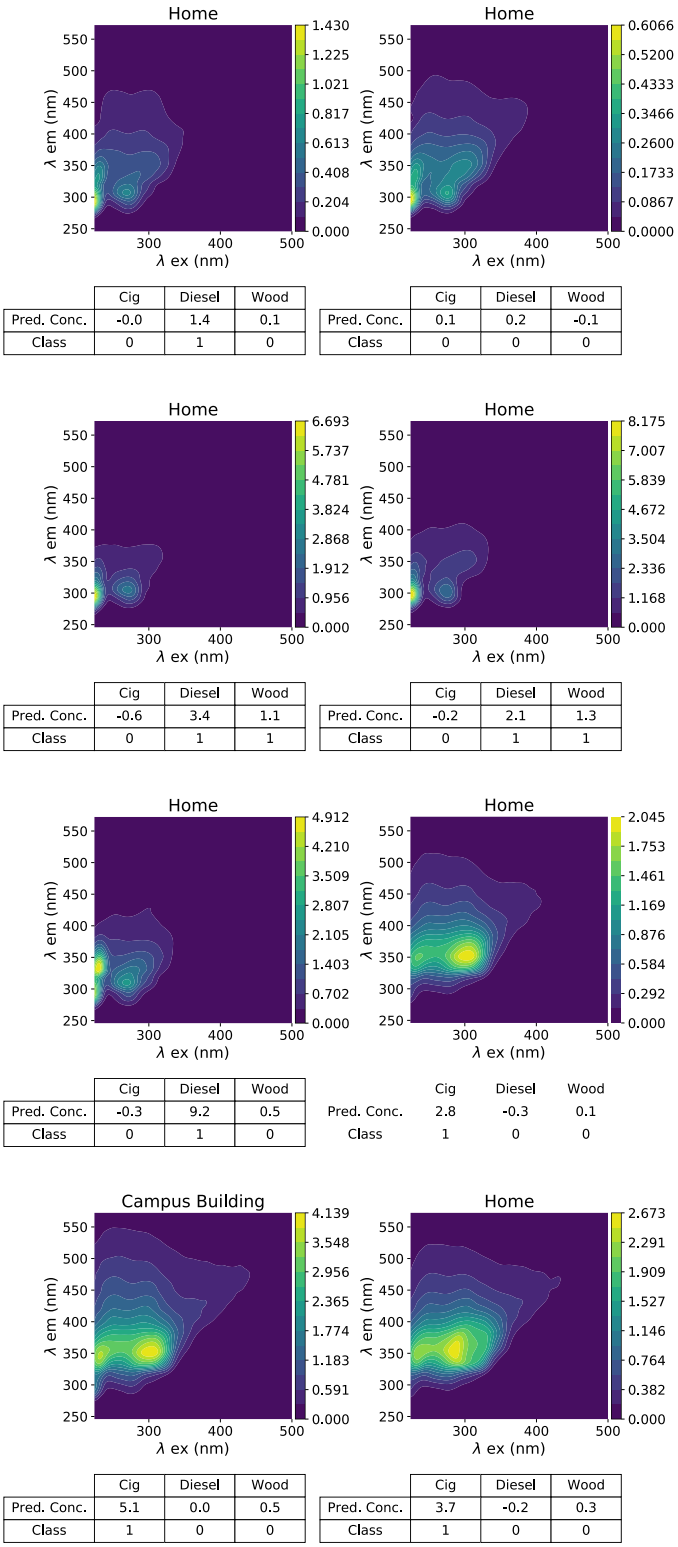
Overall Accuracy 1.00

Figure S6. Classification plots showing classification results for method blank samples. All method blank samples are classified as negative by the algorithm.

S6 Field Sample Spectra

Figure S7A shows "background" field sample spectra from Seattle homes and University of Washington buildings. The spectra are mathematically normalized to an extract concentration of 10 µg PM/mL solvent. Below each spectra a table summarizes the predicted concentration (µg PM/mL solvent) and classifies if the source is present based on the criteria applied to the laboratory samples. The classification threshold applied to the laboratory samples was 1 µg PM/mL solvent. When the normalized spectra are classified according to these criteria the algorithm will assign the source as present if the algorithm predicts the source makes up 10% or more of the spectra from the field sample.

Figure S7B shows "expected primary source" filed sample spectra collected in areas or at times when we expected the samples to be made up of a majority of cigarette, diesel or woodsmoke sources.



	Cig	Diesel	Wood
Pred. Conc.	-0.0	1.4	0.1
Class	0	1	0

	Cig	Diesel	Wood
Pred. Conc.	0.1	0.2	-0.1
Class	0	0	0

	Cig	Diesel	Wood
Pred. Conc.	-0.6	3.4	1.1
Class	0	1	1

	Cig	Diesel	Wood
Pred. Conc.	-0.2	2.1	1.3
Class	0	1	1

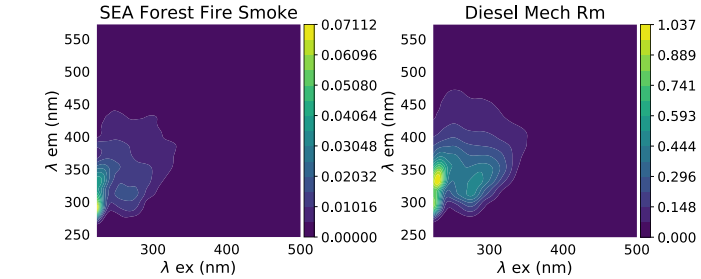
	Cig	Diesel	Wood
Pred. Conc.	-0.3	9.2	0.5
Class	0	1	0

	Cig	Diesel	Wood
Pred. Conc.	2.8	-0.3	0.1
Class	1	0	0

	Cig	Diesel	Wood
Pred. Conc.	5.1	0.0	0.5
Class	1	0	0

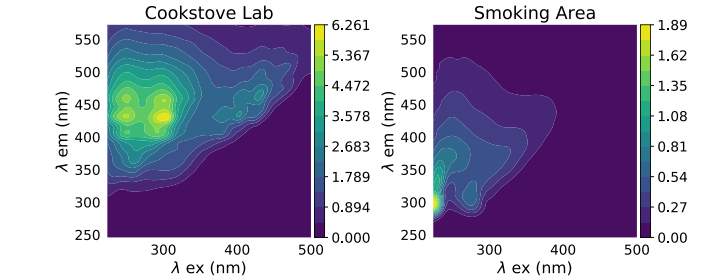
	Cig	Diesel	Wood
Pred. Conc.	3.7	-0.2	0.3
Class	1	0	0

Figure S7A. Background field samples taken in homes of researchers and inside campus buildings. The spectra look most similar to cigarette and diesel. The table below each spectrum shows the model predicted concentration of each source (µg PM/mL solvent) and the associated classification. A one indicates the spectra was classified as containing the source while a zero indicates not present.



	Cig	Diesel	Wood
Pred. Conc.	-0.0	0.2	-0.1
Class	0	0	0

	Cig	Diesel	Wood
Pred. Conc.	0.1	2.1	-0.1
Class	0	1	0



	Cig	Diesel	Wood
Pred. Conc.	-0.1	1.1	5.9
Class	0	1	1

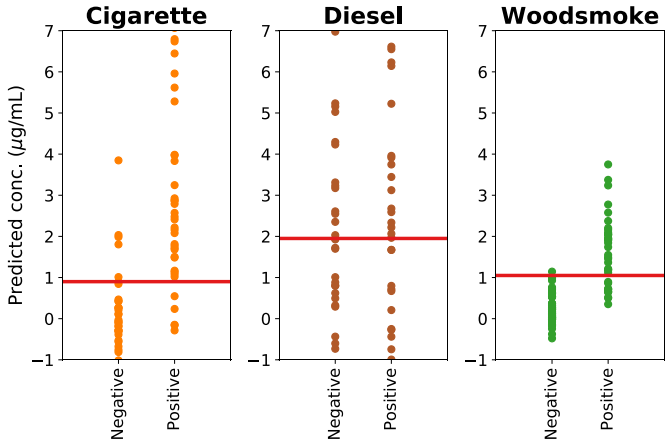
	Cig	Diesel	Wood
Pred. Conc.	0.2	2.2	0.2
Class	0	1	0

Figure S7B. Expected primary source field samples taken in locations where there was an a priori expectation that a primary source would contribute to the spectra. The table below each spectrum shows the model predicted concentration of each source (µg PM/mL solvent) and the associated classification. A one indicates the spectra was classified as containing the source while a zero indicates not present.

S7 PLS and Linear Model Results

We used the same testing and training data to assess a Partial Least Square (PLS) model. PLS, also known as "Projection to Latent Structure" is a modeling technique that projects the original predictors (fluorescent intensities) and responses (concentrations) into a lower dimensional space. The new representation of the predictors and responses are used to fit a linear model. The conceptual framework of PLS is described by G. James et al. in *An Introduction to Statistical Learning: with Applications in R* and Wegelin provides a detailed discussion of the PLS algorithm.[4,5] We implemented the PLS algorithm using Scikit-learn in Python 3.[6]

The number of dimensions in the lower dimensional space (referred to as components) is a tuning parameter for a PLS model. We fit models with 3 to 6 components. Of these models the 3-component model had the best results which are shown in Figure S8. The PLS model performance for Cigarette and Woodsmoke was acceptable with an accuracy of 88% for these sources, but it was unable to accurately identify diesel resulting in 49% accuracy for diesel and an overall accuracy of 41%.



	Cigarette	Diesel	Woodsmoke
True+	31	31	27
True-	58	18	62
False+	5	40	1
False-	7	12	11
Accuracy	0.88	0.49	0.88
Sensitivity	0.82	0.72	0.71
Specificity	0.92	0.31	0.98

Overall Accuracy 0.41

Figure S8. PLS classification results for a three-component model. The PLS model performs acceptably well on cigarette and woodsmoke but fails to classify diesel with any accuracy.

We fit a linear model to the test data set. First, we determined average EEM spectra ($\boldsymbol{S}_{avg}$) for each source using the 4 calibration spectra from each source as:

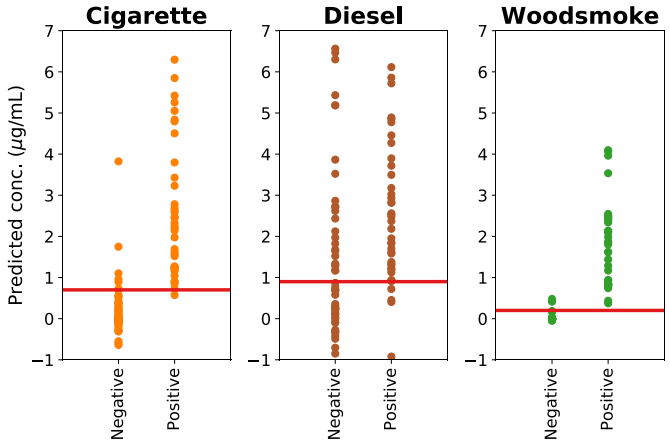
$$\boldsymbol{S}_{avg} = \sum_{i=1}^{4} \boldsymbol{S}_i / c_i,$$

where $\boldsymbol{S}_i$ is a single source calibration-spectra and c_i is the concentration of the calibrant? We assumed the 3 average spectra profiles could be used to reconstruct a test spectrum according to,

$$\boldsymbol{S} = c_C\boldsymbol{S}_C + c_D\boldsymbol{S}_D + c_W\boldsymbol{S}_W + \boldsymbol{E},$$

Where $\boldsymbol{S}$ is a test-spectra, c_C, c_d, and c_d are the predicted concentrations of cigarette, diesel and woodsmoke, respectively, $\boldsymbol{S}_C$, $\boldsymbol{S}_D$, and $\boldsymbol{S}_W$ are the average EEM spectra for cigarette, diesel and woodsmoke, respectively and E the error. To predict the three concentration values, we used the Needler-Mead optimization algorithm in SciPy to minimize the root-mean-square of $\boldsymbol{E}$.

The results of the linear model are summarized in Figure S9. The performance of this model was superior to the PLS model, with an overall accuracy of 63%, but was worse than the CNN model that had an overall accuracy of 89%. The performance of the linear model on cigarette and woodsmoke was close to the performance of the CNN with accuracies of 93% and 98% compared to 98% and 99%. The linear model performed poorly on diesel with an accuracy of only 70% compared to an accuracy of 92% for the CNN model.



	Cigarette	Diesel	Woodsmoke
True+	37	39	38
True-	57	32	61
False+	6	26	2
False-	1	4	0
Accuracy	0.93	0.70	0.98
Sensitivity	0.97	0.91	1.00
Specificity	0.90	0.55	0.97

Overall Accuracy 0.63

Figure S9. Linear model classification results. The linear model performed well on cigarette and woodsmoke but had poor performance on diesel. The poor performance on diesel led to the low overall classification accuracy of 63%

S8 Model Training

We trained the model for a total of 150 epochs. At each epoch we saved the model parameters and prediction results. At the end of the 150 epochs of training we selected the model with the best overall classification accuracy. The results presented in the paper are for a single model training run. The model was trained for 80 epochs to achieve the reported results. We repeated the training process a total of three times and each time resulted in an overall accuracy of 89%. The other two model training runs reached 89% accuracy at 73 and 85 epochs.

In Figure S10A we plot overall accuracy vs. epoch with a solid blue line and show the individual accuracies of cigarette, diesel, and woodsmoke with circles, squares, and triangles, respectively. In just a few epochs the model learns to accurately classify cigarette and woodsmoke, but it takes much longer to improve accuracy on diesel. The diesel classification accuracy varies more over each epoch than the others due the lower signal intensity of diesel and the resulting higher limit of detection.

In figure S10B we plot loss vs. epoch, using the mean-squared-error as the loss function. The training loss is noisy due to the small size of the training data set. This plot shows that the test loss reaches a minimum near 80 epochs suggesting that training beyond this point will result in overfitting.

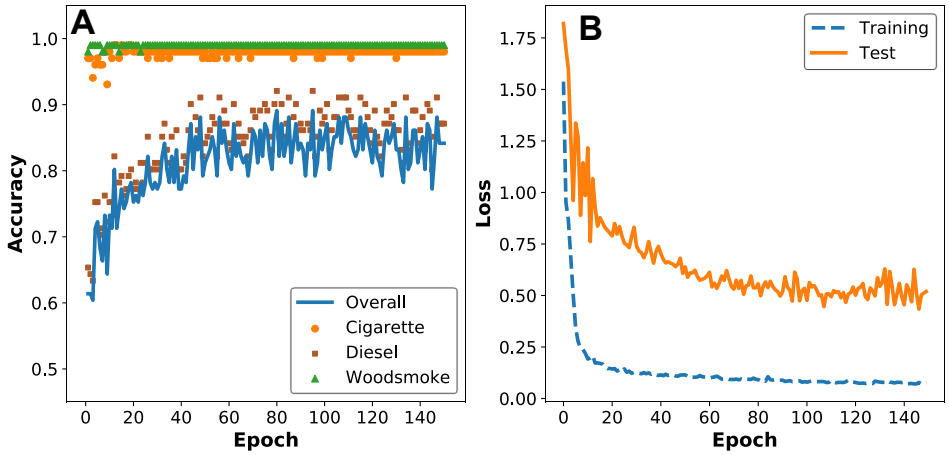


Figure S10. (A) Plot of classification accuracy as the CNN is trained for 150 epochs. Overall accuracy is shown with a solid blue line and the individual accuracies for cigarette, diesel, and woodsmoke are show with orange circles, brown squares, and green triangles, respectively. We selected the model trained for 80 epochs based on it having the highest overall classification accuracy of 89%. (B) Plot of model loss as the CNN is trained for 150 epochs. A leveling off the test loss at about 80 epochs suggests that training beyond this point results in overfitting.

S9 EEM Spectra from Water and Methanol Extracts

We evaluated extracts in various solvents by dividing filters into quarters using a scalpel. We compared spectra from extracts in ultrapure water, methanol and cyclohexane. Results were similar for methanol and cyclohexane while water extracts showed different spectra. Methanol and cyclohexane extracts were measured at lower concentrations (5ug/mL) compared to water extracts (25ug/mL and 10 ug/mL for cigarette and woodsmoke respectively). Accounting for the concentration difference between the extracts, the cyclohexane extracts had the best signal intensity. The goal of our work is to identify combustion sources of PM, so cyclohexane is a good choice for solvent given the observed signal intensity and the fact that combustion products are expected to contain non-polar fluorophores such as polycyclic aromatic compounds (PAHs).

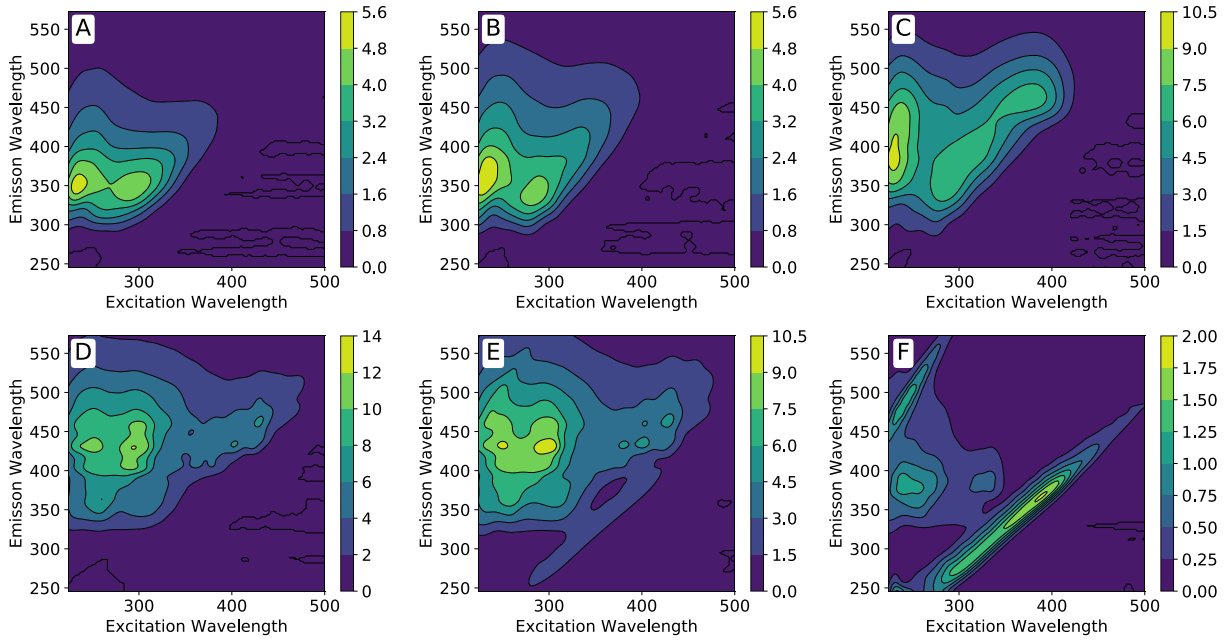


Figure S11. EEMs of in various solvents. Cigarette in cyclohexane at 5 µg/mL (A), in methanol at 5 µg/mL (B), and in ultra-pure water at 25 µg/mL (C). Woodsmoke in cyclohexane at 5 µg/mL (D), in methanol at 5 µg/mL (E), and in ultra-pure water at 10 µg/mL (F).

References

(1) Tholen, D. W. *Protocols for Determination of Limits of Detection and Limits of Quantitation: Approved Guideline*; NCCLS: Wayne, Pa., 2004.

(2) Borysiak, M. D.; Thompson, M. J.; Posner, J. D. Translating Diagnostic Assays from the Laboratory to the Clinic: Analytical and Clinical Metrics for Device Development and Evaluation. *Lab. Chip* **2016**, *16* (8), 1293–1313. https://doi.org/10.1039/c6lc00015k.

(3) Lakowicz, J. R. *Principles of Fluorescence Spectroscopy*, 3rd ed.; Springer: New York, 2006.

(4) James, G. G. M. *An Introduction to Statistical Learning: With Applications in R*; Springer texts in statistics ; 103; Springer: New York, NY, 2013.

(5) Jacob A. Wegelin. A Survey of Partial Least Squares (PLS) Methods, with Emphasis on the Two-Block Case | University of Washington Department of Statistics https://www.stat.washington.edu/index.php/article/tech-report/survey-partial-least-squares-pls-methods-emphasis-two-block-case, 2000; (accessed Dec 1, 2018).

(6) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; Duchesnay, E. Scikit-Learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.