Chemical Fingerprinting of Smoke Produced from Burning Western US Wildland Fuels

Woonyoung Chang and Ann B. Lee
Department of Statistics and Data Science, Carnegie Mellon University

Farrah Haeri and Coty Jen
Department of Chemical Engineering, Carnegie Mellon University

Dec. 2022

Abstract

Wildfires in the western US are predicted to become larger and more frequent over the next few decades due to climate change, historic fire suppression, and human encroachment into wildlands. As a result, large portions of the population are exposed to wildfire smoke as it is blown across the US. The most significant uncertainty in predicting the impact of wildfires on air quality and human health is the amounts and types of burned fuels. The goal of this project is to (i) provide interpretable representations and tools for differentiating between smoke types, and to (ii) develop methods for predicting the types and quantities of fuels present in composite smoke samples. Our data consist of chemical signatures or "fingerprints" of smoke produced from known fuels and mixtures burned in a laboratory environment. In this paper, we propose a hierarchical and geometry-based framework for visualizing and differentiating between smoke samples: a smoke sample is represented as a mixture of specific fuels, where each smoke sample and fuel is modeled as a high-dimensional distribution over chemical compounds. The distances between the compounds (the constituents of smoke) reflect the geometry of the underlying compound distribution. We show for fuel-specific burns that our approach provides interpretability and a natural means to smoothing data distributions which may reduce spurious variability in chemical fingerprints. We then illustrate how our proposed representation and metric for smoke samples can be used to determine the prevalence of specific fuels in an unknown smoke sample, and how our proposed data reduction allows the scientist to compare distributions of smoke samples via two-sample testing.

Contents

1 Introduction

Wildfires are becoming more destructive, devastating, and frequent [22, 15]. Forest managers and policymakers aim to reduce the impact of wildfires on both the regional environments and nearby communities. For example, large population areas of Los Angeles and San Francisco were blanketed in wildfire smoke for weeks each year from 2018-2021 [14]. Wildfire smoke also penetrates indoors where people spend the vast majority of their time. Consequently, millions of Americans are exposed to dangerous levels of wildfire smoke pollution. The impacts of wildfire smoke on air quality and human health are highly dependent on the amount and the type of wildland fuel burned.

Each burned fuel emits thousands of chemical compounds that define a chemical fingerprint unique to that fuel. Some fractions of these compounds are toxic and with their amounts emitted into the atmosphere are related to how much of that fuel burned and how it was burned. Therefore, understanding the link between the fuel and the specific toxic compound in smoke would allow government officials to develop effective fire and air quality models to predict the toxicity of wildfire smoke lofted to downwind communities.

Smoke modeling is a challenging problem since smoke in general consists of tens of thousands of compounds, and chemical composition depends on numerous factors such as fuel type, combustion environment, and combustion efficiency. Recent studies have been actively conducted to determine how these factors affect the composition of the compounds of smoke. [23, 7] selected ten common compounds produced from any wildland fuel. [13] used seven compounds that reported high quantitative uncertainty in the chromatographic region. These works provide useful information in differentiating smoke from ambient atmospheric samples but fall short by overlooking the thousands of remaining compounds when assessing their impact on the air quality.

In this work, we study the chemical composition of smoke produced from burning Western US Wildland fuels collected at the Blodgett Research Forest Station in Georgetown, CA. With an advanced analytic platform, consisting of ultrahigh-performance liquid chromatography (UPLC) coupled with an orbitrap mass spectrometer (OMS), we collected detailed information on the chemical composition (that is, *chemical fingerprints*) of smoke for single fuel and mixture fuel burns. See Figure 1 and Section 2.1 for details. The smoke samples are described in terms of the relative amounts of about 3,000 chemical compounds, where each compound is given by three measurements: its retention time, its parent mass, and its daughter mass spectrum (MS).

The goal of this project is to (i) provide interpretable representations and tools for differentiating between smoke types (in our case, the "single fuel burns"), and to (ii) develop methods for predicting the types and quantities of fuels present in composite smoke samples (in our case, the "mixture fuel burns"). The mass spectra convey information on the structure and chemical properties of the compounds [16]. Part of the challenge of analyzing smoke, however, lies in the high dimension (about 10,000) of these spectra. Furthermore, each smoke sample is a collection (or weighted set) of such high-dimensional spectra.

Positive matrix factorization (PMF) [17], also known as non-negative matrix factorization [12], is commonly used in chemometrics and environmental data analysis [6, 9, 5]. PMF differs from the

standard factor analysis models such as principal component analysis [3] by the constraint of positivity on both factor and loading matrices. Although PMF has been proven to be a powerful analysis tool in chemometrics, the recovered sources (loading vectors) themselves can be nonphysical and do not directly correspond to the actual sources (compounds and fuels) of smoke mixtures. In this work, we present a new approach to analyzing chemical fingerprints of smoke that directly reflects the hierarchical structure of compounds, fuels and fuel mixtures. Our representation of smoke captures the underlying structure of chemical compounds and directly links observed smoke to a group of compounds with common characteristics.

There are three key parts to our approach: (1) defining a histogram representation of smoke samples that provides interpretability, (2) defining a distance metric on smoke samples that takes the inter-bin relationship of their histogram representations into account, and finally (3) leveraging the proposed histogram representation and associated distance metric in quantification (determining the prevalence of fuels in a mixture fuel) and two-sample testing (comparing distributions of smoke samples).

Hence, as a starting point, we represent the smoke sample as histograms where the bins correspond to groups of compounds. This quantization is a form of smoothing, allowing a trade-off between bias and variance in density estimation; it also serves as a geometry-based compression of the data. The desired properties of the partitioning are that (i) similar compounds fall into the same bins, and that (ii) it provides us with a principled means to quantifying the similarity between compounds as well as between bins (groups of compounds). In this work, we borrow the diffusion-based approach proposed in [11] to learn the geometry of the underlying distribution of the data. The diffusion framework provides a new coordinate system, where we embed bins as well as smoke samples, and where the Euclidean distance reflects the local geometry of the data. The proposed distributional diffusion distance (DDD) is induced by the diffusion map of compounds and the quantification; in fact, the DDD is defined as the Euclidean distance between embedded smoke samples in the coarse-grained diffusion space. In our work, we investigate the optimal bin selection based on the classification performance of single-fuel burns.

The induced DDD metric has close connections with the well-known statistical distances between probability distribution such as the earth mover's distance (EMD) [18] and the maximum mean discrepancy (MMD) [4]. EMD computes the optimal transport cost of moving one distribution to another based on a pre-defined ground distance between the bins of the histogram. For DDD, the cost to move mass is naturally set as the distance between embedded bins. Whereas EMD often suffers from its high computation cost, DDD has a linear time complexity [11]. We also review the MMD, a distance between distributions based on embedding probabilities in a reproducing kernel Hilbert space. We show in Remark 1 that when we treat each compound as an individual bin of smoke histogram, DDD is the same as MMD. Hence, DDD can be viewed as a generalization of MMD but with the advantage of being able to control the degree of compression and smoothing via partitioning in compound space.

Equipped with a histogram representation of smoke samples, we present a method for determining the prevalence of a specific fuel in an unknown smoke sample. This method has been adapted from the machine learning literature on solving classification under prior probability shift [20]. Finally, we introduce a two-sample test for comparing two classes of smoke distributions. The proposed two-sample testing algorithm is fully non-parametric and involves classifying each smoke sample into one of two groups based on the DDD metric.

The rest of the paper is outlined as follows: We begin by describing the experimental design and the notation in Section 2. Then, in Section 3.1, we introduce a proposed histogram representation of smoke samples and define the statistical distance between histograms. Equipped with the representation, we describe the quantification method in Section 3.2 and study the two-sample test in Section 3.3. In Section 4.2 and 4.2.2, we demonstrate the interpretability of the proposed representation of smoke samples and confirm that smoke samples are well-classified with the induced metric. We then apply our methods for determining the prevalence of specific fuel in the smoke sample and comparing two classes of smoke distribution in Section 4.3.

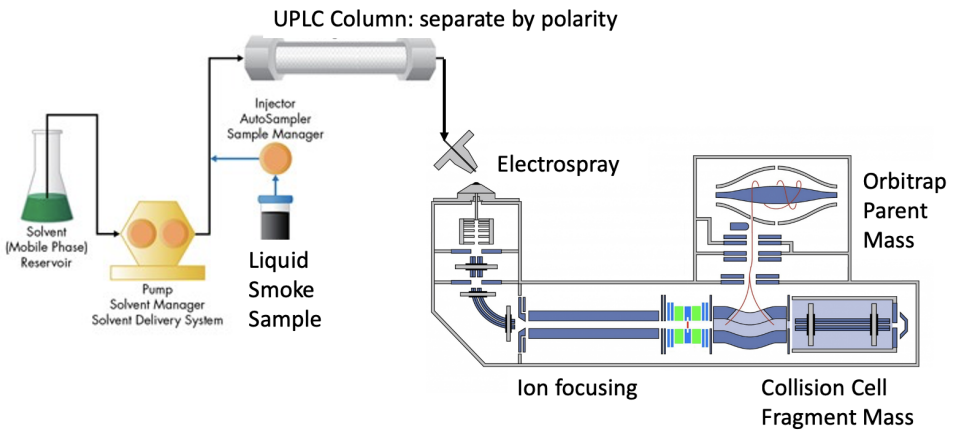


Figure 1: Molecular speciation by ultra-high performance liquid chromatography (UPLC) orbitrap mass spectrometer.

2 Chemical Fingerprinting

2.1 Experimental Set-Up

Smoke samples were analyzed by ultrahigh-performance liquid chromatography (UPLC) coupled to an orbitrap mass spectrometer (OMS). Figure 1 schematically illustrates the process of identifying the chemical compounds present in the smoke sample. The smoke was first dissolved into a liquid solvent before being injected into the UPLC. The UPLC consists of a C18 column which separates molecules by their polarity and outputs a chromatogram of the smoke sample as in Figure 2a. More polar compounds elute from the column faster with a shorter so-called *retention time* than less polar compounds. The separated compounds are then ionized by electrospray ionization, and the ions travel to the OMS to have their unfragmented so-called *parent mass* measured. The compound is then fragmented into smaller daughter ions in the collision dissociation chamber. These fragment ions finally travel back to the OMS to have the masses and intensities of the daughter ions measured, resulting in a *daughter mass spectrum* which is a normalized histogram plot of the intensity versus the mass-to-charge ratio. An example daughter mass spectrum can be seen in Figure 2b.

2.2 Data Collection

Smoke was produced by burning fuels in the CMU Air Quality Laboratory's burn chamber. Fuels were collected from the Blodgett Research Forest Station located in Georgetown, CA. Three types of dead fuels are analyzed here and include duff, litter, and woody debris. These fuels are common on the forest floor in the Western US. Each fuel was individually burned five times, and we also burned mixture fuels having the following mass ratios*; 5:5:0, 7:3:0, and 1:1:1. These combination fuel burns were repeated three times, and three to four smoke samples were collected at different points in time for each burn. In total, 43 fuel-specific smoke samples and 39 mixture smoke samples were collected.

2.3 Data Representation and Notation

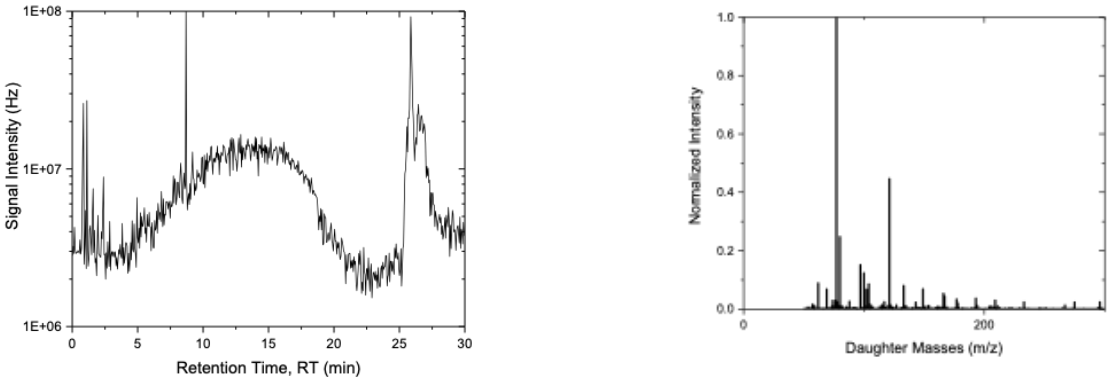
We here elaborate on the notation that we use in the rest of the paper.

Let a smoke sample be

$$\mathcal{D} = \left(\left\{(\mathbf{x}_1, w_{\mathbf{x}_1}), \ldots, (\mathbf{x}_p, w_{\mathbf{x}_p})\right\}, f\right) \tag{1}$$

where $\mathbf{x}_1, \ldots, \mathbf{x}_p$ ($p \approx 3,000$) are chemical compounds found in the smoke sample and the weights $w_{\mathbf{x}_1}, \ldots, w_{\mathbf{x}_p}$ are the corresponding normalized emission intensities with $\sum_{i=1}^{p} w_{\mathbf{x}_i} = 1$. The fuel label f indicates which fuel induces the given smoke sample. Compounds are identified via three measurements: two scalar variables, *retention time* and *parent mass*, and a high-dimensional *daughter mass spectrum*. This triplet defines the chemical fingerprint of an individual compound. More precisely,

*Each ratio corresponds to the contents of litter, duff, and woody debris in order.



(a) (b)

Figure 2: (a) A chromatogram of a smoke sample, where each peak corresponds to a chemical compound with a unique daughter mass spectrum. In the chromatogram, the height of the peak on the y-axis corresponds to the "signal intensity" or the proportion of the compound in the smoke sample; the location on the x-axis represents the compound's measured "retention time" in the UPLC column. (b) The measured daughter mass spectrum of compound $C_4H_6O_4$ consists of the set of signal intensities of daughter ions forming the compound, where each daughter ion is identified by its mass-to-charge ratio.

the chemical fingerprint is given by $\mathbf{x}_i = \{t_i, m_i, \mathbf{v}_{\mathrm{MS},i}\}$ where $t_i \in \mathbb{R}$, $m_i \in \mathbb{R}$, and $\mathbf{v}_{\mathrm{MS},i} \in \mathbb{S}^{q-1}$ ($q \approx 10,000$), respectively, denote the retention time, the parent mass, and the daughter mass spectrum of compound $\mathbf{x}_i$. The mass spectrum $\mathbf{v}_{\mathrm{MS}}$ is represented as a unit vector in order to use the structural information of ions, rather than the peak heights at specific ions.

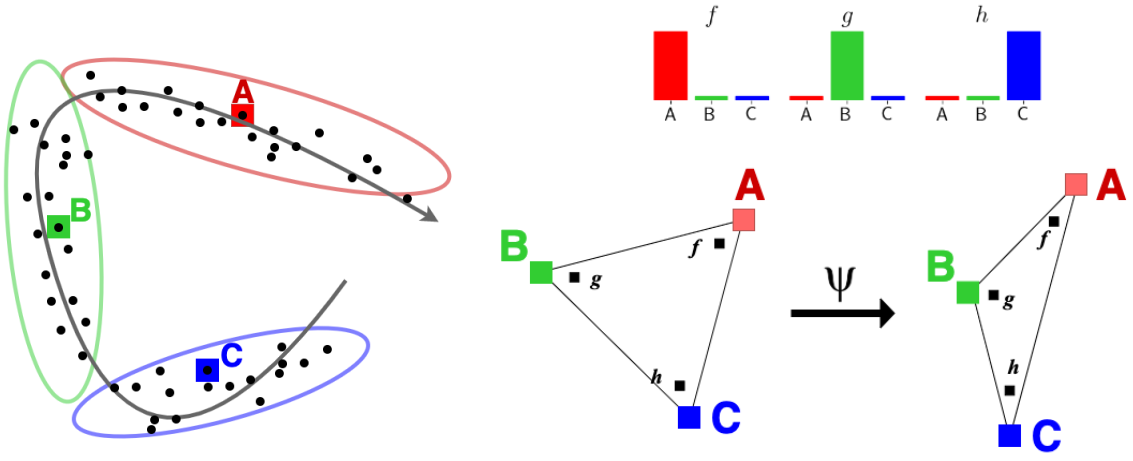
In this paper, we investigate two categories of smoke: (i) fuel-specific smoke from pure fuel incineration and (ii) mixture smokes from mixture fuels with known composition. We denote the probability density of the fuel-specific smoke which is induced by a specific fuel f over the compounds be $p(\mathbf{x}|f)$. Also, we write a general smoke from fuel mixture burn as $p(\mathbf{x}|s)$.

3 Methods

3.1 Distributional Diffusion Distance (DDD): Defining a Distance between Smoke Samples

Diffusion maps are nonlinear dimension reduction techniques for problems where linear methods such as principal component analysis (PCA) may fail. Through diffusion maps, we aim to build a coordinate system that preserves the local geometry of data points. In our application, the compounds lie in a high-dimensional space, but the data display a low-dimensional structure due to physicochemical constraints.

The schematic illustration in Figure 3 conveys the key idea of our methodology. Suppose the data possess a one-dimensional intrinsic structure as depicted in (a). The intrinsic geometry of the data suggests that the data points in cluster A are closer to the points in cluster B than those in cluster C. The top panel in (b) displays three smoke samples as histograms over these regions, and note that we can correspond histograms to points in a simplex. According to the intrinsic structure of data, f is closer to g than to h, but the canonical metric in Cartesian coordinates may dictate that f, g, and h are at *equal* distances from each other as can be seen in the left bottom panel in the left triangle of the panel (b). In contrast, the right triangle of the panel (b) depicts the embedded compound groups and smoke samples in the diffusion space, and their relative locations take the intrinsic structure of data into account. Furthermore, we can naturally define a distance between histograms as a Euclidean distance between points in the diffusion space. Below we start by describing the *diffusion distance*, a geometry-based distance between individual compounds. This distance defines a *diffusion map* to a Euclidean space, where we introduce the *distributional diffusion distance*, a new metric for distributions



(a) (b)

Figure 3: Schematic diagram of the key idea behind our geometry-based metric. (a) Data (chemical compounds) along a low-dimensional manifold. (b) Three smoke samples (distributions over compounds) represented as coarse-grained histograms over three groups of compounds (top). The Cartesian coordinates do not capture the intrinsic data structure, positioning f, g, and h at equal distances (bottom left). A diffusion map gives a better representation of the geodesic distances between both compounds and smoke samples (bottom right).

over compounds.

Diffusion Distance. To capture the local geometry of data, we begin by constructing a weighted graph. Let $\mathcal{X}$ be a collection of compounds and let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a similarity function. A common kernel in spectral graph methods is the Gaussian kernel,

$$k(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{\rho(\mathbf{x}, \mathbf{x}')^2}{\sigma^2}\right), \quad \mathbf{x}, \mathbf{x}' \in \mathcal{X},$$

where $\rho(\cdot, \cdot)$ is a distance between nodes. In our particular application, we employ the cosine distance which is widely used to compare mass spectra [19, 8],

$$\rho(\mathbf{x}, \mathbf{x}') = 1 - \frac{\mathbf{v}_{\mathrm{MS}} \circ \mathbf{v}'_{\mathrm{MS}}}{\|\mathbf{v}_{\mathrm{MS}}\| \, \|\mathbf{v}'_{\mathrm{MS}}\|},$$

where $\mathbf{v}_{\mathrm{MS}}$ is a normalized daughter mass spectrum of compound $\mathbf{x}$, and $\circ$ denotes a dot product. The bandwidth σ determines the effective neighborhood of compounds, with smaller values of σ leading to less overlap between distant compounds. See Section A.4 for more details on the choice of the kernel function.

A weighted graph $(\mathcal{X}, k)$ is a collection of nodes (compounds) $\mathcal{X}$ and weights k on edges that capture the similarity between two nodes. Let the kernel matrix $\mathbf{K}$ of a weighted graph $(\mathcal{X}, k)$ be a $p \times p$ matrix $\mathbf{K} = (k(\mathbf{x}_i, \mathbf{x}_j))_{i,j=1}^{p}$, and let the degree matrix $\mathbf{D} = \mathrm{Diag}(d_i)_{i=1}^{p}$ be a $p \times p$ diagonal matrix whose entries are the degrees of nodes in the graph; that is, $d_i = \sum_j k(\mathbf{x}_i, \mathbf{x}_j)$. We then define a row-wise normalized graph Laplacian,

$$\mathbf{A} = \mathbf{D}^{-1}\mathbf{K}.$$

The matrix $\mathbf{A}$ often serves as a main tool in spectral clustering. See [21] and the reference therein for a thorough discussion of the graph Laplacian.

In diffusion maps, we define a Markov Chain on the graph where the 1-step transition probability matrix is given by the matrix $\mathbf{A}$. This chain integrates the local transition probabilities between compounds, giving higher weights to higher-density regions in compound space. The t-step diffusion distance between compounds $\mathbf{x}$ and $\mathbf{x}'$ is defined as a weighted L^2 distance between t-step transition kernels,

$$D^t(\mathbf{x}, \mathbf{x}') = \left\| \mathbf{A}^t(\mathbf{x}, \cdot) - \mathbf{A}^t(\mathbf{x}', \cdot) \right\|_{1/\phi_0}, \tag{2}$$

Algorithm 1 K-Means Algorithm

Require: set of compounds $\mathcal{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_p\}$; kernel k; embedding dimension d; number of clusters K

Calculate the adjacency matrix $\mathbf{K}$ and the degree matrix $\mathbf{D}$.
Construct the transition matrix $\mathbf{L} = \mathbf{D}^{-1}\mathbf{K}$.
Compute the first d eigenvalues of $\mathbf{L}$ and the associated right-eigenvectors.
Construct the averaged diffusion map defined in (4).
Randomly assign $\mathbf{x}_1, \ldots, \mathbf{x}_p$ to K clusters.
repeat
    Compute the geometric centroid of each cluster defined in (6).
    Reassign each $\mathbf{x}_1, \ldots, \mathbf{x}_p$ to the cluster which has the closest centroid.
until convergence
return $(\mathcal{C}_i, \mathbf{c}_i)_{i=1}^K$

where ϕ_0 is the stationary density of the Markov Chain.

Diffusion Map. The diffusion distance in (2) can be directly computed using the spectral decomposition of the transition probability matrix $\mathbf{A}$. Let $1 = \lambda_0 \geq \lambda_1 \geq \ldots \geq \lambda_d$ be the leading d eigenvalues of $\mathbf{A}$ in decreasing order having $\mathbf{1}_p = \psi_0, \psi_1, \ldots, \psi_d$ as right-eigenfunctions. We define a t-step diffusion map Ψ^t along the path connecting any two compounds;

$$\Psi^t(x) = (\lambda_1^t \psi_1(x), \ldots, \lambda_d^t \psi_d(x))^\top. \tag{3}$$

Instead of a fixed t, we use an averaged diffusion map combining diffusion at all times t,

$$\overline{\Psi}(x) = \left(\frac{\lambda_1}{1-\lambda_1}\psi_1(x), \ldots, \frac{\lambda_d}{1-\lambda_d}\psi_d(x) \right)^\top. \tag{4}$$

It is shown in [11] that the L^2-distance in the new coordinate system approximates the diffusion distance,

$$\left\| \Psi^t(\mathbf{x}) - \Psi^t(\mathbf{x}') \right\| \approx D(\mathbf{x}, \mathbf{x}'). \tag{5}$$

Diffusion map in (3) or (4) provides a new coordinate system that takes the local geometry of the data point into account. The quality of the truncated approximation in (5) depends on the embedding dimension d. In general, the embedding dimension d is chosen by inspecting the degree of eigenvalue drop; we choose $d = 100$ for our analysis.

Distributional Diffusion Distance. Equipped with the diffusion embedding, we next quantize the compound space $\mathcal{X}$ to allow us to efficiently estimate the marginal density $p(\mathbf{x})$ over the compound space. Using the quantized representation, we estimate the conditional density $p(\mathbf{x}|f)$ — the distribution of smoke for fuel f. More specifically, we partition the compound space according to the diffusion K-means algorithm described in Algorithm 1. The algorithm outputs K clusters $\{\mathcal{C}_i : i = 1, \ldots, K\}$ (where each chemical compound is assigned to one cluster based on diffusion distances), together with the geometric centroid $\mathbf{c}_i$ or associated coordinate of each cluster $\mathcal{C}_i$ in diffusion space,

$$\mathbf{c}_i = \mathbf{c}(\mathcal{C}_i) = \frac{\sum_{j:\mathbf{x}_j \in \mathcal{C}_i} d_j \overline{\Psi}(\mathbf{x}_j)}{\sum_{j:\mathbf{x}_j \in \mathcal{C}_i} d_j}, \quad i = 1, \ldots K. \tag{6}$$

From a theoretical perspective, the solution of the diffusion K-means algorithm in 1 approximately preserves the spectral property through an associated coarse-grained Markov Chain on the K centroids [10].

We next represent the smoke sample $\mathcal{D} = \left(\left\{ (\mathbf{x}_1, w_{\mathbf{x}_1}), \ldots, (\mathbf{x}_p, w_{\mathbf{x}_p}) \right\}, f \right)$ as a K-bin histogram, which has its support on K diffusion clusters $\mathcal{C}_1, \ldots, \mathcal{C}_K$;

$$\widetilde{p}(\mathbf{x}|f) = \sum_{i=1}^{K} w_{\mathcal{C}_i} \mathbb{1}(\mathbf{x} \in \mathcal{C}_i) \tag{7}$$

where $\mathbb{1}$ denotes a binary indicator function and

$$w_{\mathcal{C}_i} = \sum_{j:\mathbf{x}_j \in \mathcal{C}_i} w_{\mathbf{x}_j}, \quad i = 1, \ldots, K.$$

The geometric centroid represents the relative location of each cluster or bin in diffusion space. It follows that the smoke sample in (7) corresponds to the point $\mathbf{w}$ within the $(K-1)$-simplex having K vertices at geometric centroids $\mathbf{c}_1, \ldots, \mathbf{c}_K$ in the d-dimensional diffusion space,

$$\mathbf{w} = w_{\mathcal{C}_1}\mathbf{c}_1 + \ldots + w_{\mathcal{C}_K}\mathbf{c}_K. \tag{8}$$

If, for instance, a particular smoke contains a large proportion of compounds belonging to cluster $\mathcal{C}_i$, then the point $\mathbf{w}$ in (8) would be close to the centroid $\mathbf{c}_i$. See the schematic diagram in Figure 3 for a similar example where, e.g., the smoke sample (histogram) f in panel (b) has a large proportion of compounds in bin A, and is consequently represented by a point close to the representative centroid for cluster A in the diffusion map to the bottom right of Figure 3b.

Finally, we define the *distributional diffusion distance* (DDD) between two smoke samples $\mathcal{D}$ and $\mathcal{D}'$ as

$$\widehat{\mathrm{DDD}}\left(\widetilde{p}(\mathbf{x}|f), \widetilde{p}(\mathbf{x}|f')\right) = \left\|\mathbf{w} - \mathbf{w}'\right\|_2, \tag{9}$$

where $\widetilde{p}(\mathbf{x}|f)$ and $\widetilde{p}(\mathbf{x}|f')$ are K-bin representations of smoke samples $\mathcal{D}$ and $\mathcal{D}'$, respectively, and $\mathbf{w}$ and $\mathbf{w}'$ represent their locations in the d-dimensional diffusion space.

As evident in the construction, the DDD metric reflects the often nonlinear geometry of the data. Furthermore, the DDD metric provides interpretability, because it not only provides a distance between two conditional distributions but also allows within-bin comparison between two distributions.

3.2 Quantification: Evaluating the Prevalence of a Given Fuel in a Smoke Sample

For a smoke sample s, let $p(\mathbf{x}|s)$ be the K-bin histogram representation as in (7). Denote a set of base fuels as $\{f_1, \ldots, f_F\}$, then we have from Bayes theorem that

$$p(\mathbf{x}|s) = \sum_{i=1}^{F} p(\mathbf{x}|f_i, s)p(f_i|s).$$

Here, the conditional probability $p(f_i|s)$ can be interpreted as the proportion of a specific fuel f_i in a fuel composite s. Our goal in this section is to introduce a statistical methodology that allows us to recover the fuel composition from mixture smoke; i.e., to estimate $p(f_i|s)$. The general methodology for solving this type of problem has previously been developed for a different machine learning problem, namely that of prior probability shift [20] in transfer learning. The key assumption that allows learning $p(f_i|s)$ is the prior probability shift assumption,

$$\mathbf{x} \perp\!\!\!\perp s|f_i, \quad i = 1, \ldots, F.$$

In our context, the condition means that the fuel-conditional distribution of the smoke over the compounds, $p(\mathbf{x}|f_i, s)$, is invariant to the composition of s. That is, under this assumption, the distribution of smoke from the fuel mixture s can be written as

$$p(\mathbf{x}|s) = \sum_{i=1}^{F} \alpha_{f_i} p(\mathbf{x}|f_i), \quad \sum_{i=1}^{F} \alpha_{f_i} = 1, \tag{10}$$

where α_{f_i} denotes the unknown proportion of fuel f_i in the fuel composite s.

Now consider a fixed real-valued function g. From (10), we have that

$$\mathbb{E}[g(\mathbf{x})|s] = \sum_{i=1}^{F} \alpha_{f_i} \mathbb{E}[(g(\mathbf{x})|f_i].$$

Given K clusters of compounds, $\mathcal{C}_k (k = 1, \ldots, K)$, note that $\mathbb{E}[g(\mathbf{x})|f]$ can be computed as

$$\mathbb{E}[g(\mathbf{x})|f] = \sum_{k=1}^{K} g(\mathcal{C}_k)\mathbb{P}(\mathbf{x} \in \mathcal{C}_k|f),$$

for $f \in \{f_i : i = 1, \ldots, F\} \cup \{s\}$. Here $\mathbb{P}(\mathbf{x} \in \mathcal{C}_k|f)$ can be estimated via

$$\widehat{\mathbb{P}}(\mathbf{x} \in \mathcal{C}_k|f) = \sum_{\mathbf{x}_i \in \mathcal{C}_k} \tilde{w}^f_{\mathbf{x}_i},$$

where $\tilde{w}^f_{\mathbf{x}_i}$ denotes the measured proportion of compound $\mathbf{x}_i$ in smoke generated by fuel f. Therefore, an estimate of coefficients α_{f_i} can be obtained by solving a linear equation with $F - 1$ unknown parameters,

$$\widehat{\mathbb{E}}[g(\mathbf{x})|s] = \sum_{i=1}^{F} \alpha_{f_i} \widehat{\mathbb{E}}[(g(\mathbf{x})|f_i],$$

where one degree of freedom is missing due to the sum-to-one constraint. Therefore, with $(F - 1)$ different choices of g, our estimate $\hat{\boldsymbol{\alpha}} = (\widehat{\alpha}_{f_1}, \ldots, \widehat{\alpha}_{f_F})^\top$ solves a linear system of equations,

$$\mathbf{g}_s = \mathbf{G}\boldsymbol{\alpha}$$

where $\mathbf{G} = (\mathbf{G}_{i,j})_{i,j=1}^F$ such that $G_{i,j} = \widehat{\mathbb{E}}[g_i(\mathbf{x})|f_j]$ and $\mathbf{g}_s = (\widehat{\mathbb{E}}[g_1(\mathbf{x})|s], \ldots, \widehat{\mathbb{E}}[g_F(\mathbf{x})|s])^\top$ with $g_1 \equiv 1$. In principle, any choices of g's yield an estimator whereas the quality of the estimator, on the other hand, depends on the choice. One natural choice is $g_i(\mathcal{C}_k) = \widehat{\mathbb{P}}(f|\mathbf{x} \in \mathcal{C}_k) \propto \widehat{\mathbb{P}}(\mathbf{x} \in \mathcal{C}_k|f_i)$ for $i = 2, \ldots, F$. An induced estimator is known to have approximate minimax property [20]. Since $\widehat{\boldsymbol{\alpha}}$ may have negative coefficients, we use the projected estimator $\widehat{\boldsymbol{\beta}}$ of $\widehat{\boldsymbol{\alpha}}$, which is, $\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}:\sum_{i=1}^F \beta_i = 1, \beta_i \geq 0} \|\boldsymbol{\beta} - \widehat{\boldsymbol{\alpha}}\|_2$.

3.3 Two-Sample Test: Comparing Distributions of Smoke Samples

The general objective of a two-sample test is to determine whether two classes are the same or not based on observed data. For our application, our goal is to test

$$H_0 : p(\mathbf{x}|Y = 0) = p(\mathbf{x}|Y = 1) \tag{11}$$

where the conditional distribution $p(\mathbf{x}|Y)$ is the distribution of smoke over compounds $\mathbf{x}$ and the label $Y \in \{0, 1\}$ indicates the class of the compound. With K-bin representations of smoke samples in (7), testing (11) is, by Bayes theorem, equivalent to testing

$$H_0 : \mathbb{P}(Y = 1|\mathbf{x} = x) = \mathbb{P}(Y = 1) \quad \text{for all } x \in \mathcal{X}. \tag{12}$$

Hence, we compare the regression function $\mathbb{P}(Y = 1|\mathbf{x} = x)$ with the class probability $\mathbb{P}(Y = 1)$. Specifically, we consider the following discrepancy between the two distributions,

$$T = \int_{\mathcal{X}} [\mathbb{P}(Y = 1|\mathbf{x} = x) - \mathbb{P}(Y = 1)]^2 \, dp_{\mathcal{X}}(\mathbf{x} = x). \tag{13}$$

When estimating $\mathbb{P}(Y = 1|\mathbf{x} = x)$, we can take advantage of regression methods that adapt to the structure of data. Here we first represent smoke samples as K-bin histograms in a d-dimensional diffusion space, then train an n-nearest neighbor (NN) regressor based on the approximated DDD, and finally estimate (13) by computing the discrepancy measure at a set of pre-defined evaluation points. That is, we define the test statistic

$$\widehat{T} = \sum_{x \in \mathcal{V}} (\widehat{m}(x) - \widehat{\pi})^2, \tag{14}$$

where $\mathcal{V}$ is a set of evaluation points, $\widehat{m}$ is an estimate of $\mathbb{P}(Y = 1|\mathbf{x} = x)$, and $\widehat{\pi}$ is an estimate of $\mathbb{P}(Y = 1)$. The null distribution of the proposed statistic in (14) is generally unknown as it depends on the choice of regression method and evaluation points. However, one can perform a permutation or Monte Carlo (MC) test, yielding a p-value from an empirical distribution of the test statistic.

4 Results

4.1 Explanatory Data Analysis (EDA)

First, we explore how the defining features of an individual compound (e.g., a triplet of retention time, parent mass, and daughter mass spectrum) vary between compounds. This analysis is done due to non-idealities in how the UPLC-OMS operates. For example, two different compounds with almost identical retention times will enter the collision cell at the same time due to the time resolution of the instrument. The resulting convoluted daughter mass spectrum will contain ion fragments from both compounds. As a result. each daughter mass spectrum needed to be compared to the parent mass to ensure the spectrum represented a single compound. According to the heuristic matching criteria described in Appendix A.1, we ended up using only 2,905 compounds (out of 3,359 observed compounds) and their chemical fingerprints in our analysis.

4.1.1 Relationship between Parent Mass and Retention Time

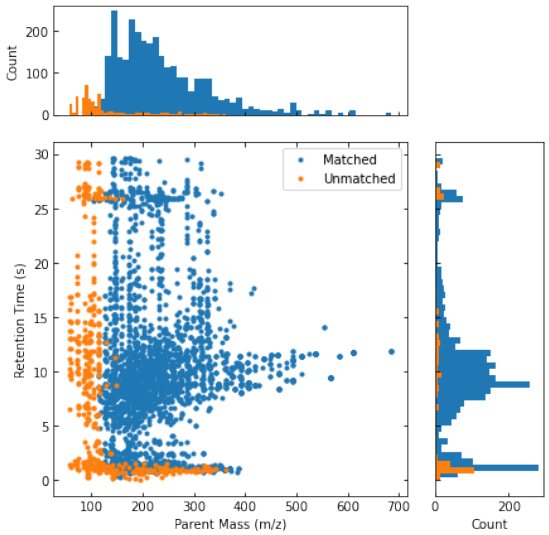
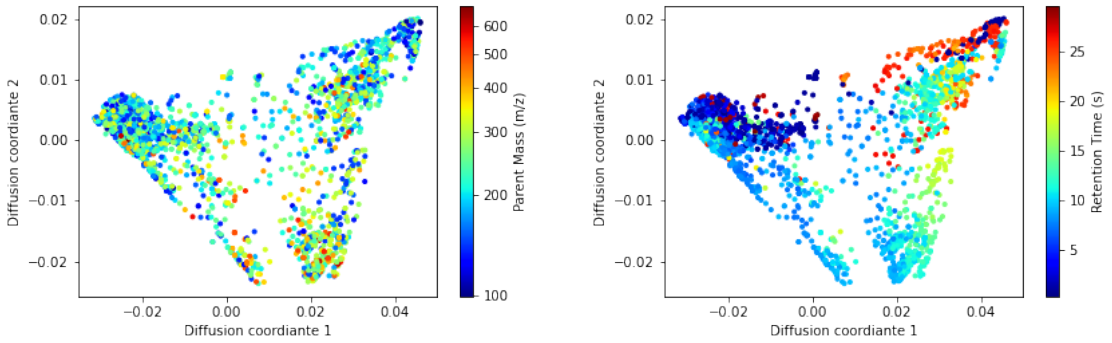


Figure 4: Scatter plot of the measured retention time and parent mass for the 3359 compounds found in single-fuel and mixture-fuel burns. Unmatched compounds (shown in orange color) tend to have either small masses or short retention time. Only the 2,905 matched compounds (shown in blue color) are used in our analysis.

Figure 4 shows a scatter plot of the measured retention time t_j and parent mass m_j for all $j = 1, \ldots, 3359$ compounds observed in our single-fuel and mixture-fuel burns. The compounds that did not match the mass spectra according to our matching algorithm are shown in orange color, and the remaining $2,905$ compounds are shown in blue color. Unmatched compounds either have a small parent mass or a small retention time which suggests these unmatched compounds are background noise of the instrument. We exclude the unmatched compounds from our analysis since such degeneracy may prevent compounds from being well-identified.

We note that the distribution of parent mass is skewed towards large mass values, whereas the distribution of retention time is skewed towards small retention times. In addition, the retention times of a majority of compounds tend to concentrate on 0, indicating that the vast majority of compounds are highly polar. A closer inspection of the mass axis reveals a degeneracy in the parent mass. That is, we observe sets of compounds with similar parent mass, although they clearly correspond to different compounds with distinct retention times. We speculate that the observed degeneracy is due to isomers — molecules with identical formulae (hence, identical parent mass) but with different arrangements of atoms in the molecule, leading to different chemical or physical properties such as polarity which results in different retention times. Hence, distinguishing between different chemical compounds requires using not only parent mass but also retention time or mass spectrum information.

4.1.2 Structure in Daughter Mass Spectra as Revealed by Diffusion Maps



(a) (b)

Figure 5: Each compound in the left and right panels is colored by its measured (a) parent mass and (b) retention time, respectively. We see from the right panel that diffusion maps based on daughter mass spectra are recovering the information on retention times, although this information was not directly included in the construction of diffusion maps.

The EDA in Section 4.1 showed that parent mass alone cannot distinguish between different compounds or, more specifically, between isomers. Hence, we will base the rest of our analysis on the measured daughter mass spectra. These spectra are distinct for different chemical compounds. Unlike retention times, daughter mass spectra can also be used to determine the chemical identity and structure of molecules.

As a starting point, we construct the diffusion map based on the cosine similarity between the daughter mass spectra of two compounds, as described in Section 3.1. Figure 5 depicts a two-dimensional visualization of a diffusion map of chemical compounds colored by parent mass or retention time. Panel (b) shows that the diffusion map recovers the retention time of compounds, with compounds having similar retention times falling into nearby regions in the diffusion space. This is despite the fact that the retention time measurements were not explicitly used in the construction of the diffusion map.

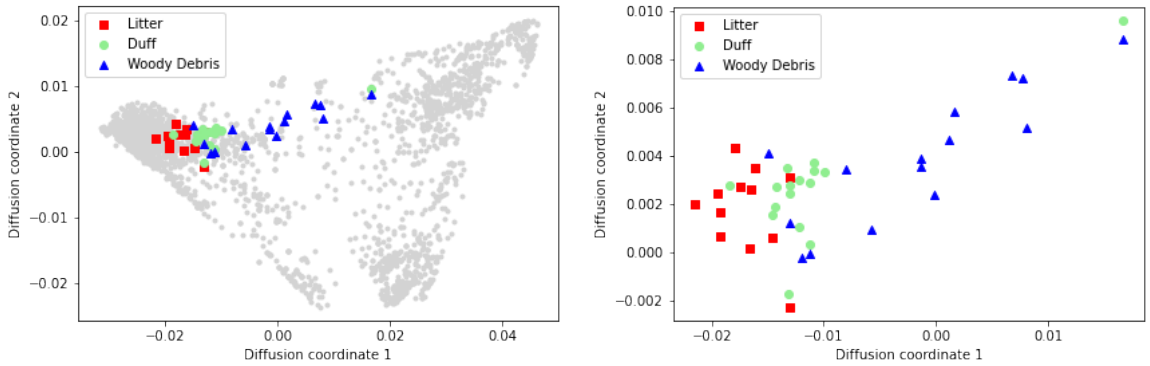
4.2 Differentiating Smoke Samples via Distributional Diffusion Distance

In this section, we evaluate whether the distributional diffusion distance (DDD) metric defined in Section 3.1 produces a reasonable separation of fuel-specific samples. It is noteworthy to mention that the chemical composition of smoke depends on a number of variables, including the combustion efficiency describing the stage of the burn (such as the initial flaming and final smoldering) and environmental factors describing the conditions during the experiment. However, in this work, we assume that smoke produced by burning the same type of fuel will be close in terms of our chosen metric, despite not taking into account differences due to e.g. the stage of the burn.

Recall from Section 3.1 that we represent smoke samples as K-bin histograms over groups of compounds (7); these histograms correspond to points in a $(K-1)$-simplex with the K vertices of the simplex corresponding to the locations of the bins (8) in diffusion space. Figure 14 displays the diffusion map results for $K = 2,905$ — the case where we treat each compound as a separate bin of the histogram, and each vertex corresponds to one vertex of the simplex. We observe that the DDD without coarse-graining produces a clear grouping of smoke samples into three types of fuel. In the following sections, we show that these results can be further improved by coarse-graining the diffusion.

4.2.1 Diffusion Maps Provide Interpretability

The general idea of the diffusion K-means procedure described in Algorithm 1 is to construct K partitions of the compound space. This induces a smoothed representation of smoke samples in terms of K-bin histograms with a geometry-based distance between different bins.



(a) (b)

Figure 6: Two-dimensional visualization of diffusion map of chemical compounds (gray dots) and 43 fuel-specific smoke samples from duff (green circle), litter (red square), and woody debris (blue triangle).

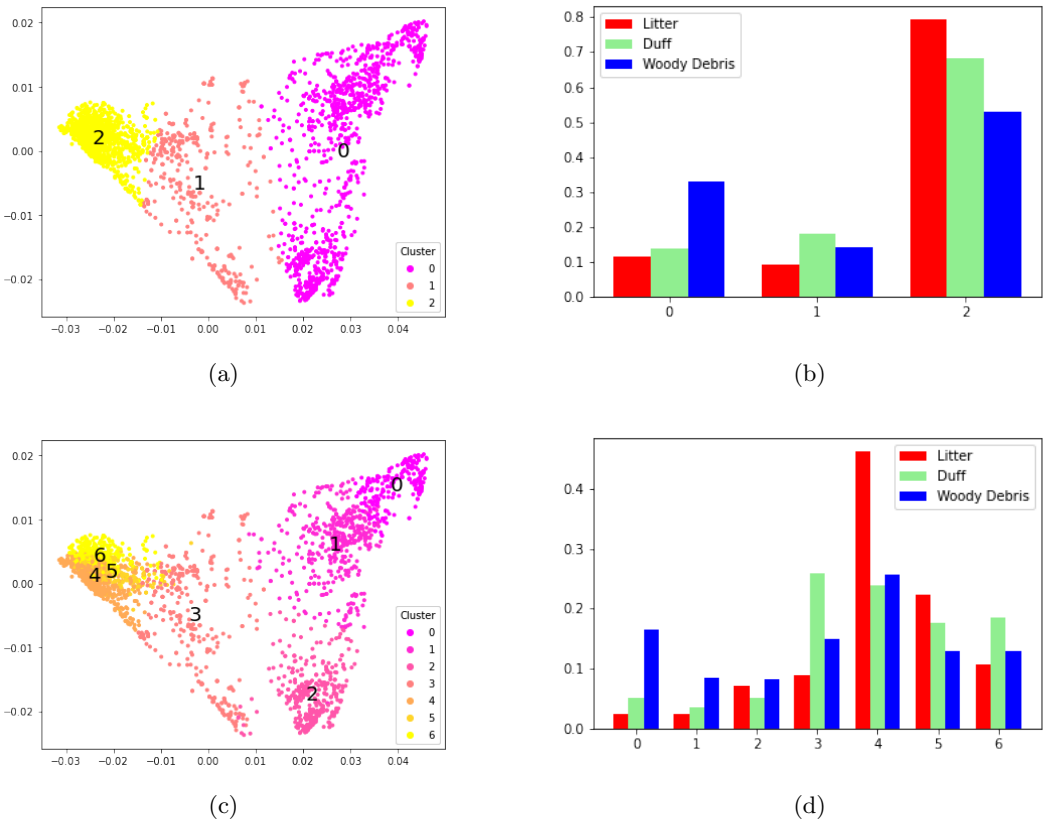
The smallest value of K that would convey the information in a two-dimensional arrangement would be $K = 3$. Figure 7a and 7b show a minimal representation of compounds and the smoke samples in the diffusion space with $K = 3$, respectively. Even a 3-bin representation conveys meaningful information. Figure 7b shows that the majority of compounds from the three fuel-specific burns fall in Cluster 2, which is located in the left corner of the diffusion map in Figure 7a. Nevertheless, we expect that a choice of $K = 3$ would represent "over-smoothing", as we are compressing the information from 2,905 compounds into that of only three representative compound clusters.

A larger number of bins provides more information on the detailed structure of the smoke samples. Figure 7c and 7d show the results for $K = 7$. The left corner of the triangle – where the large proportion of compounds are located – is now divided into Clusters 4, 5, and 6. It is also noteworthy that for example, the litter tends to generate more compounds in cluster 4 compared to other fuels when burned. On the other hand, we can see that burning the woody debris releases compounds in Cluster 0 while the duff releases compounds corresponding to Cluster 3. Finer partition of compounds may provide a more precise link between specific fuel types and groups of compounds. However, a certain amount of smoothing may be necessary due to spurious noise in the data. Later, in Section 4.2.2, we search for the optimal choice of K based on the degree of separation of the smoke samples by the types of burned fuel.

4.2.2 Diffusion Coarse-Graining Provides Geometry-Based Compression

The DDD depends on the choice of K yielding different relative locations of smoke samples in the diffusion space. To objectively determine the right amount of smoothing or the number of partitions, K, we implement receiver operating characteristic (ROC) analysis based on the n-nearest-neighbor regression. For a fixed n, we fit the n-nearest classifier in the 100-dimensional embedding of diffusion space under 5-fold cross-validation. Then we compute the true positive rate and the false positive rate of each fitted classifier. As our application deals with a three-way classification problem, we use the "one versus rest" (OvR) method [2]. The OvR method compares each class against all other classes and gives 3 ROC curves as shown in Figure 8a. By averaging three curves, we obtain a final OvR ROC curve; its corresponding OvR area under the curve (AUC) values is depicted in Figure 8b. When K is small, the performance improves rapidly as the number of bins increases. When $K \geq 10$, there is no notable improvement. As discussed in Section 3.1 and Remark 1, the DDD converges to MMD as the number of bins increases. Hence, as expected, the DDD classification performance approaches the performance of MMD samples as K increases. We can obtain comparable or even better results using, for instance, only $K = 7$ bins. This is a meaningful computation gain considering the linear time complexity of our method.

In Figure 9, we compare the visualizations of fuel-specific smoke samples for three different choices of K. Overall, the results show a fair level of separation between fuel types considering that we are only viewing two-dimensional projections of the data. In particular, the choice of $K = 7$ yields the



(a) (b)

(c) (d)

Figure 7: The results of the diffusion K-means algorithm for two different choices of K. The *top row* shows (a) the diffusion map for $K = 3$ and (b) the induced distribution of fuel-specific smokes in compound space for $K = 3$. The *bottom row* shows the (c) diffusion map for $K = 7$ and (d) the induced distribution of fuel-specific smokes for $K = 7$.

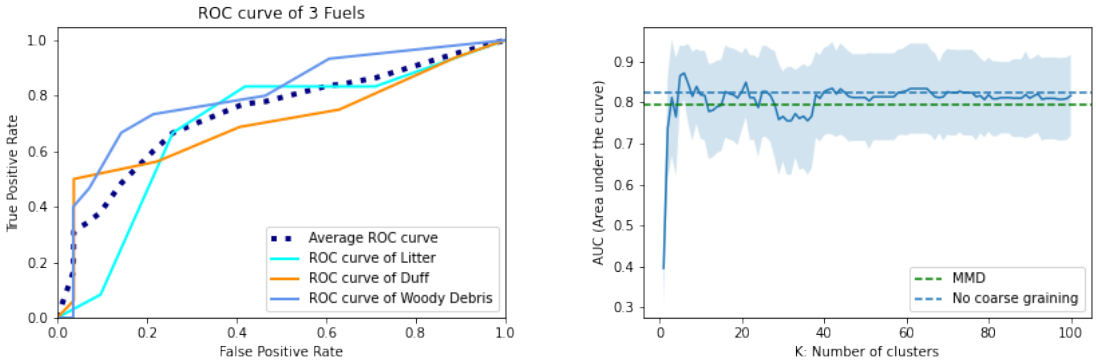
right amount of smoothing, achieving the best classification for our data. Hence, we will represent smoke samples with $K = 7$ bin histograms throughout the rest of the paper.

4.3 Quantification and Goodness-of-Fit Test for Smoke Samples

We start this section by applying the quantification method in Section 3.2 to the 39 smoke samples from the mixture fuel burns. Figure 10, however, shows that the estimated weights tend to be far from the true weights. In this section, we investigate whether the poor results are due to the quantification process itself, or due to the assumptions behind the procedure being violated.

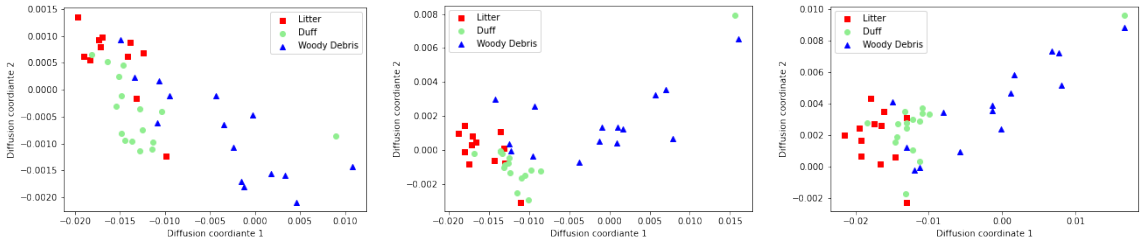
Recall that the key assumption for the quantification method in Section 3.2 is the prior probability shift assumption. In our application, this assumption means that the chemical composition of a specific fuel does not depend on other fuels that occur in the same sample. Hence, if fuels react during a burn, the prior probability shift assumption may no longer be a reasonable assumption.

In what follows, we apply the same quantification method to synthetic mixture smoke, which are artificially constructed to satisfy the prior probability shift assumption by linearly mixing representations of single-fuel lab burns. The generation process is described in Algorithm 2. Figure 10 shows the quantification results of the synthetic mixture samples. Although the estimated weights show a fair amount of uncertainty, the overall results for the synthetic mixture are more reasonable than those of the real fuel mixture. First, this suggests that our quantification method can roughly recover the true proportion of fuels under the prior probability shift assumption. Second, it raises the question: Is the smoke from the fuel mixture a linear mixture of fuel-specific smoke? In what follows, we use a goodness-of-fit Monte Carlo test to answer whether the lab mixtures are significantly different from the synthetic linear mixtures. This test is a version of the two-sample test described in Section 3.3.



(a) (b)

Figure 8: Classification performance of fuel-specific samples using a n-nearest neighbor classifier and the DDD metric with varying levels of coarse-graining, K. Panel (a) corresponds to the example case $K = 2,905$, where we treat all compounds as individual bins. The three solid lines represent one-versus-rest ROC curves for the three classes duff, litter, and woody debris. Averaging the three curves results in the "Average ROC curve" (dotted). Panel (b) shows the classification performance for varying K, where the performance is measured in terms of the area under the average ROC curve. The blue ribbon indicates the magnitude of the CV standard error of the AUC estimate. The two horizontal lines correspond to the AUC value for the case $K = 2,905$ (purple dashed) and a two-dimensional MDS map (described in Section A.3) based on pairwise MMD between smoke samples (green dashed). The gap between the two lines is due to the small embedding dimension of the MDS map (2 versus 100 for diffusion maps).



(a) $K = 3$ (b) $K = 7$ (c) $K = 2,905$

Figure 9: Visualizations of smoke samples in diffusion space for (a) $K = 3$ minimal (b) $K = 7$ best and (c) $K = 2,905$ no coarse-graining.

Panel (a) and (b) of Figure 12 show how the synthetic smoke mixtures (blue markers) versus the real mixture fuel burns (orange markers) are distributed in diffusion space. However, we may not able to detect the difference in a two-dimensional visualization as the smoke samples are lying in a (7-1)-simplex. To detect whether there is a statistically significant difference in distribution between synthetic versus real smoke, we perform the two-sample test described in Section 3.3. Our goal is to test $\mathbb{P}(\mathbf{x}|Y = 0) = \mathbb{P}(\mathbf{x}|Y = 1)$, where the label $Y = 0$ and $Y = 1$ indicate a synthetic mixture and real fuel mixture, respectively. We use the real mixture sample for training and randomly draw samples from both synthetic mixture and smoke samples from the fuel mixture to form the evaluation set. Finally, test statistics are based on a five-nearest neighbor regression on $d = 100$-dimensional diffusion space. The procedure is described in Algorithm 3. Panel (c) shows the distribution of our test statistic under the null hypothesis that two classes of smoke, synthetic mixture and real fuel mixture, are equal in distribution. All tests fail to accept the null indicating that the synthetic mixture smoke and real smoke from the fuel mixture have differences in their distributions for all three types of mixtures.

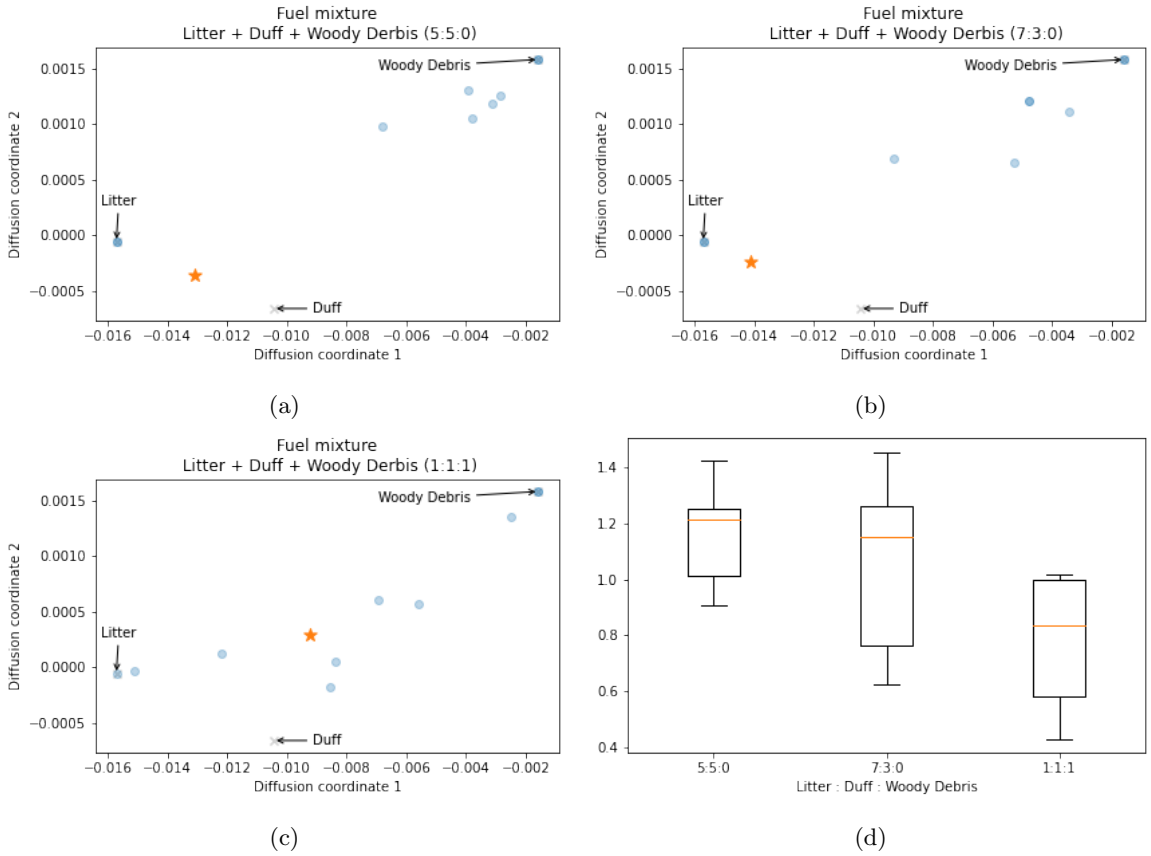


Figure 10: Quantification results for smoke samples from mixture fuel burns with (a) 5:5 mixture of litter and duff, (b) 7:3 mixture of litter and duff, and (c) 1:1:1 mixture of litter, duff, and woody debris. We reconstructed the smoke samples using the estimated weights. The reconstructed smoke samples are shown as blue dots on a two-simplex, where the vertices here are the centers of fuel-specific smoke samples. An orange star corresponds to a reconstruction with the true fuel proportions. (d) Distributions of the ℓ_2 loss $\|\boldsymbol{\alpha} - \hat{\boldsymbol{\alpha}}\|_2$ of the estimated weights for each mixture fuel burn.

Algorithm 2 Synthetic Mixture Generation

Require: a set of K-bin representations $p(\mathbf{x}|f)$ of fuel-specific smoke; set $\mathcal{S}$ fuel-specific smoke; labels of pure fuels $\mathcal{F} = \{f_1, \ldots, f_F\}$; size of synthetic mixture samples M; desired fuel composition of synthetic mixture $\boldsymbol{\alpha} \in \mathbb{R}^F$

Ensure: a set of K-bin representations $p(\mathbf{x}|s)$ of synthetic mixture smoke

1: **for** $m \in \{1, \ldots, M\}$ **do**
2: **for** $i \in \{1, \ldots, F\}$ **do**
3: Sample a histogram $p(x|f_i) = \sum_{k=1}^{K} w_{\mathcal{C}_k}^i \mathbb{1}(\mathbf{x} \in \mathcal{C}_k)$ given a fuel f_i
4: **end for**
5: Construct a mixture histogram as the linear combination of fuel-specific histograms; $\tilde{p}^{(m)}(\mathbf{x}|s) = \sum_{k=1}^{K} \sum_{i=1}^{F} \alpha_i w_{\mathcal{C}_k}^i \mathbb{1}(\mathbf{x} \in \mathcal{C}_k)$.
6: Perturb each bin of $\tilde{p}^{(m)}(\mathbf{x}|s)$ with a small amount; $\tilde{p}^{(m)}(\mathbf{x} \in \mathcal{C}_k|s) \leftarrow \tilde{p}^{(m)}(\mathbf{x} \in \mathcal{C}_k|s) + \epsilon_k$
7: Renormalize $\tilde{p}^{(m)}(\mathbf{x}|s)$
8: **end for**
9: **return** $\{\tilde{p}^{(m)}(\mathbf{x}|s) : m = 1, \ldots, M\}$

5 Conclusions

In this paper, we developed a novel representation of the chemical composition of smoke with each smoke sample being represented as a K-bin histogram over groups of compounds. This representa-

Algorithm 3 Goodness-of-Fit Regression Test via Monte Carlo (MC) Sampling

Require: train set $\mathcal{S}$ of N measurements of mixture fuel (denote the distribution of such measurements by F); synthetic model F_0 for simulating linear mixtures; the size of Monte Carlo sample M where $M \gg N$; the number of additional Monte Carlo samples B; a regression method (such as kNN based on DDD?) for estimating $m_{\text{post}}(\mathbf{x}) := \mathbb{P}(Y = 1 \mid \mathbf{X} = \mathbf{x})$; set of evaluation points $\mathcal{V}$

Ensure: p-value for testing if $H_0 : F = F_0$ for all $\mathbf{x} \in \mathcal{V}$; local posterior differences $\{\lambda(\mathbf{x})\}_{\mathbf{x}\in\mathcal{V}}$ at the evaluation points $\mathbf{x} \in \mathcal{V}$

`Compute test statistic at the evaluation points` $\mathbf{x} \in \mathcal{V}$
Let $n_{\text{tot}} = N + M$.
Sample $\mathcal{S}^* = \{\mathbf{X}_1^*, \ldots, \mathbf{X}_M^*\}$ from F_0.
Define an augmented sample $\{\mathbf{X}_i, Y_i\}_{i=1}^{n_{\text{tot}}}$, where $\{\mathbf{X}_i\}_{i=1}^{n_{\text{tot}}} = \mathcal{S} \cup \mathcal{S}^*$, and $Y_i = I(\mathbf{X}_i \in \mathcal{S})$.
Estimate $m_{\text{prior}} := \mathbb{P}(Y = 1)$ with class proportion $\widehat{m}_{\text{prior}} = N/n_{\text{tot}}$.
Compute $\widehat{m}_{\text{post}}$ using augmented sample.
Compute test statistic $\lambda = \sum_{\mathbf{x}\in\mathcal{V}} \lambda^2(\mathbf{x})$, where $\lambda(\mathbf{x}) = \widehat{m}_{\text{post}}(\mathbf{x}) - \widehat{m}_{\text{prior}}$
`Estimate the null distribution of the test statistic`
for $b \in \{1, \ldots, B\}$ **do**
    Sample $\mathcal{S}^{(b)} = \{\mathbf{X}_1^{(b)}, \ldots, \mathbf{X}_N^{(b)}\}$ from F, under the null hypothesis $H_0 : F = F_0$.
    Sample $\mathcal{S}^{*(b)} = \{\mathbf{X}_1^{*(b)}, \ldots, \mathbf{X}_M^{*(b)}\}$ from F_0.
    Define a new augmented MC sample $\{\mathbf{X}_i, Y_i\}_{i=1}^{n_{\text{tot}}}$, where $\{\mathbf{X}_i\}_{i=1}^{n} = \mathcal{S}^{(b)} \cup \mathcal{S}^{*(b)}$, and $Y_i = I(\mathbf{X}_i \in \mathcal{S}^{(b)})$.
    Refit $\widehat{m}_{\text{post}}$ and calculate the test statistic on the new augmented sample to obtain $\widetilde{\lambda}^{(b)}$ from its distribution under $H_0 : F = F_0$.
end for
`Compute approximate` p`-value`
Compute the p-value by $\widehat{p} = \frac{1}{B+1}\left(1 + \sum_{b=1}^{B} I(\widetilde{\lambda}^{(b)} > \lambda)\right)$.
return $\widehat{p}$, $\{\lambda(\mathbf{x})\}_{\mathbf{x}\in\mathcal{V}}$

tion takes the relationship between the chemical signatures of the compounds into account and links differences in smoke composition to groups of compounds. A K-bin representation further aids interpretability as one can directly visualize smoke samples in diffusion space as points in the convex hull of the K representative compounds. The proposed metric (DDD), induced by the histogram representation, reflects the local geometry of the data. It represents a de-noised metric through data smoothing. We can control the degree of smoothing by adjusting the number of bins K, and we found that the choice of $K = 7$ bins yielded better separation of smoke samples than using the full set of $N \sim 3,000$ compounds. With the histogram representation of the smoke sample, we then presented a new quantification method that estimates the unknown proportion of fuels in the smoke generated by burning fuel mixtures. The key assumption is "prior probability shift", which here means that the composite smoke is a linear mixture of fuel-specific smoke. Lastly, we adopted a fully non-parametric two-sample test to detect statistically significant differences between two classes of smoke, such as actually observed smoke versus output from a simulation model. We found statistical evidence indicating that smoke from our fuel mixtures is not a linear mixture of fuel-specific smoke. That is, the key assumption of prior probability shift behind our quantification method may not hold for this particular application.

References

[1] Michael A. A. Cox and Trevor F. Cox. *Multidimensional Scaling*, pages 315–347. Springer Berlin Heidelberg, Berlin, Heidelberg, 2008.

[2] André CPLF de Carvalho and Alex A Freitas. A tutorial on multi-label classification techniques. *Foundations of computational intelligence volume 5*, pages 177–195, 2009.

[3] Karl Pearson F.R.S. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572, 1901.

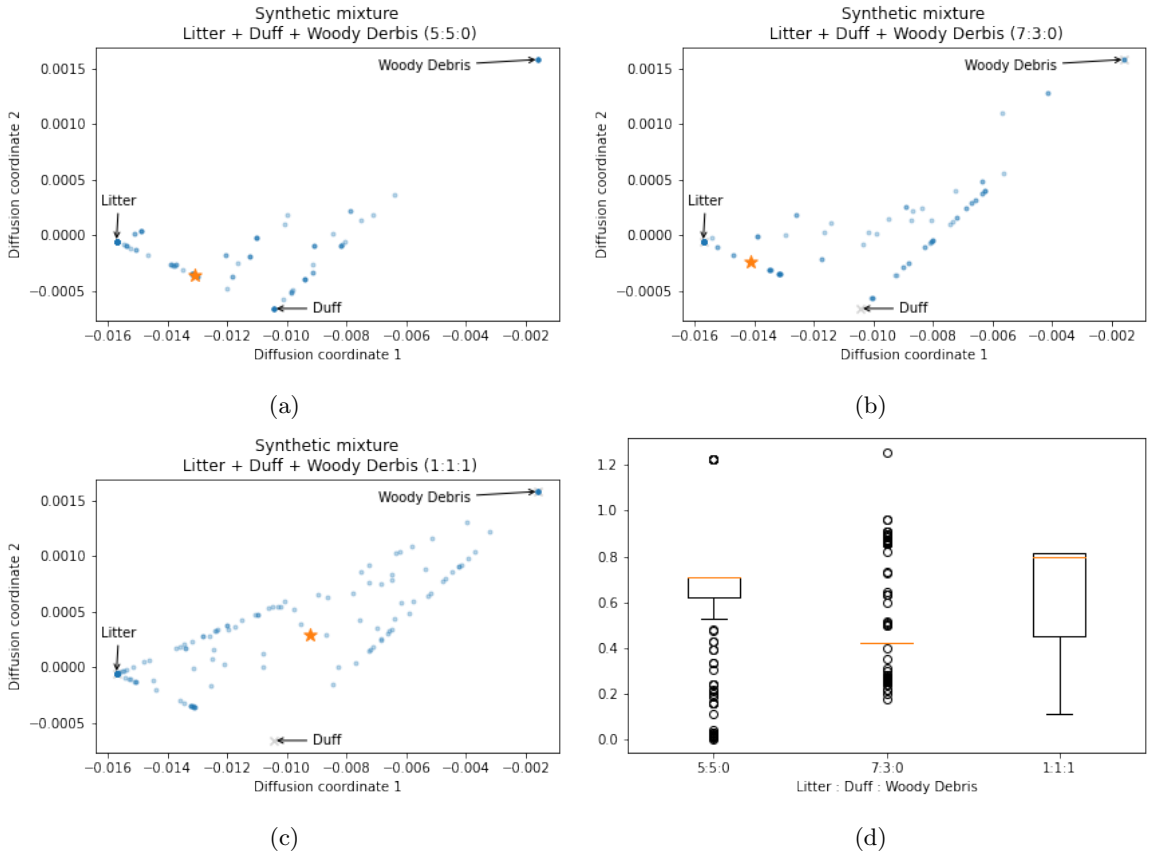


Figure 11: Quantification results for synthetic smoke samples generated by Algorithm 2 with (a) 5:5 mixture of litter and duff, (b) 7:3 mixture of litter and duff, and (c) 1:1:1 mixture of litter, duff, and woody debris. The reconstructed smoke samples are shown as blue dots as in Figure 10. (d) Distributions of the ℓ_2 loss of the estimated weights for each synthetic mixture.

[4] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773, 2012.

[5] Indrani Gupta, Abhaysinh Salunkhe, and Rakesh Kumar. Source apportionment of pm10 by positive matrix factorization in urban area of mumbai, india. *The Scientific World Journal*, 2012, 2012.

[6] Wen-Yuan Ho, Kuo-Hsin Tseng, Ming-Lone Liou, Chang-Chuan Chan, and Chia-hung Wang. Application of positive matrix factorization in the identification of the sources of pm2. 5 in taipei city. *International journal of environmental research and public health*, 15(7):1305, 2018.

[7] Sharon Hood and Duncan Lutes. Predicting post-fire tree mortality for 12 western us conifers using the first order fire effects model (fofem). *Fire Ecology*, 13(2):66–84, 2017.

[8] Florian Huber, Lars Ridder, Stefan Verhoeven, Jurriaan H Spaaks, Faruk Diblen, Simon Rogers, and Justin JJ Van Der Hooft. Spec2vec: Improved mass spectral similarity scoring through learning of structural relationships. *PLoS computational biology*, 17(2):e1008724, 2021.

[9] Abigail R Koss, Manjula R Canagaratna, Alexander Zaytsev, Jordan E Krechmer, Martin Breitenlechner, Kevin J Nihill, Christopher Y Lim, James C Rowe, Joseph R Roscioli, Frank N Keutsch, et al. Dimensionality-reduction techniques for complex mass spectrometric datasets: application to laboratory atmospheric organic oxidation experiments. *Atmospheric chemistry and physics*, 20(2):1021–1041, 2020.

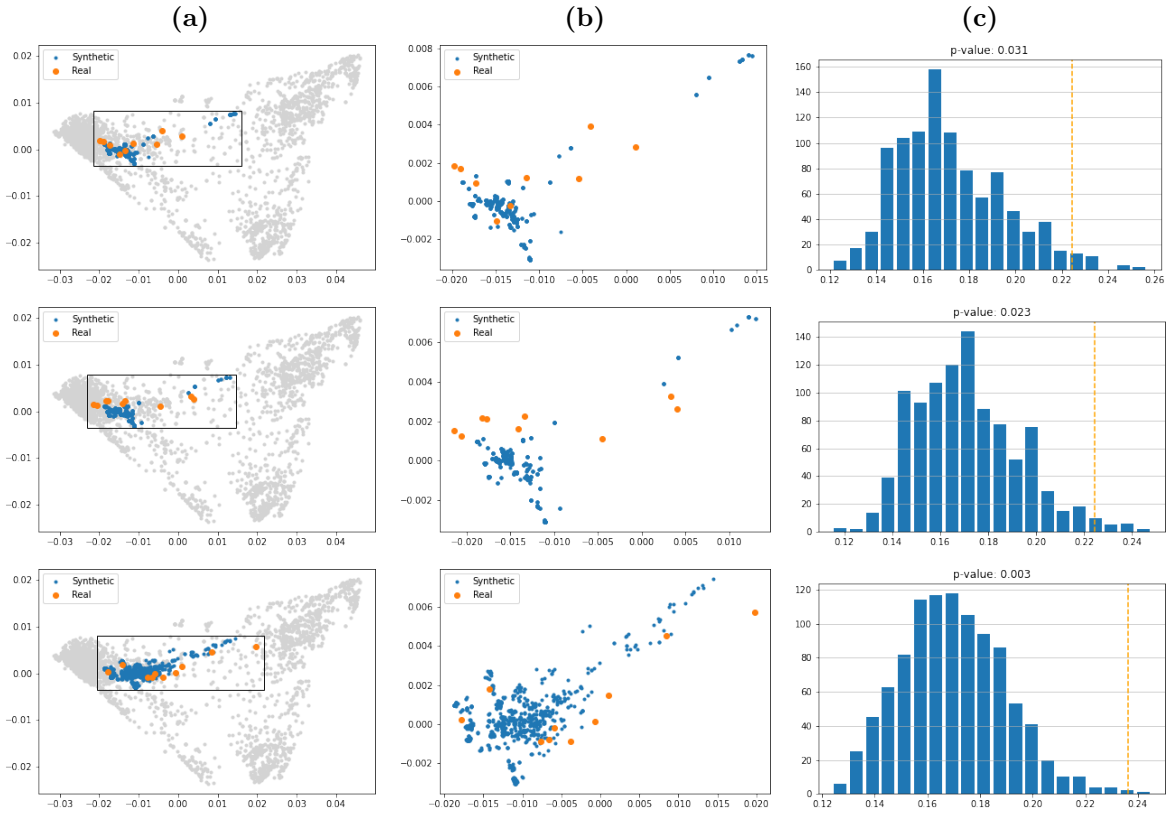


Figure 12: (a)(b) Smoke samples for synthetic mixtures (blue) and fuel-mixture lab burns (orange) in diffusion space for 5:5 mixture of litter and duff (row 1), 7:3 mixture of litter and duff (row 2), and 1:1:1 mixture of litter, duff, and woody debris (row 3). (c) Distribution of the test statistic (14) under the null that two classes of smoke, the synthetic mixtures and the lab fuel mixtures, have the same distributions where the vertical lines indicate the observed values of the test statistic. For all three fuel mixtures, the p-values indicate that the two classes have significantly different distributions.

[10] S. Lafon and A.B. Lee. Diffusion maps and coarse-graining: a unified framework for dimensionality reduction, graph partitioning, and data set parameterization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(9):1393–1403, 2006.

[11] Stéphane S Lafon. *Diffusion maps and geometric harmonics*. Yale University, 2004.

[12] Daniel Lee and H Sebastian Seung. Algorithms for non-negative matrix factorization. *Advances in neural information processing systems*, 13, 2000.

[13] Yutong Liang, Coty N Jen, Robert J Weber, Pawel K Misztal, and Allen H Goldstein. Chemical composition of pm 2.5 in october 2017 northern california wildfire plumes. *Atmospheric Chemistry and Physics*, 21(7):5719–5737, 2021.

[14] Yutong Liang, Deep Sengupta, Mark J Campmier, David M Lunderberg, Joshua S Apte, and Allen H Goldstein. Wildfire smoke impacts on indoor air quality assessed using crowdsourced data in california. *Proceedings of the National Academy of Sciences*, 118(36):e2106478118, 2021.

[15] Jennifer R Marlon, Patrick J Bartlein, Daniel G Gavin, Colin J Long, R Scott Anderson, Christy E Briles, Kendrick J Brown, Daniele Colombaroli, Douglas J Hallett, Mitchell J Power, et al. Long-term perspective on wildfires in the western usa. *Proceedings of the National Academy of Sciences*, 109(9):E535–E543, 2012.

[16] Fred W McLafferty. Mass spectrometry across the sciences. *Proceedings of the National Academy of Sciences*, 105(47):18088–18089, 2008.

[17] Pentti Paatero and Unto Tapper. Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics*, 5(2):111–126, 1994.

[18] Yossi Rubner, Carlo Tomasi, and Leonidas J Guibas. The earth mover's distance as a metric for image retrieval. *International journal of computer vision*, 40(2):99–121, 2000.

[19] Stephen E Stein and Donald R Scott. Optimization and testing of mass spectral library search algorithms for compound identification. *Journal of the American Society for Mass Spectrometry*, 5(9):859–866, 1994.

[20] Afonso Fernandes Vaz, Rafael Izbicki, and Rafael Bassi Stern. Quantification under prior probability shift: The ratio estimator and its extensions. *The Journal of Machine Learning Research*, 20(1):2921–2953, 2019.

[21] Ulrike Von Luxburg. A tutorial on spectral clustering. *Statistics and computing*, 17(4):395–416, 2007.

[22] Anthony LeRoy Westerling. Increasing western us forest wildfire activity: sensitivity to changes in the timing of spring. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 371(1696):20150178, 2016.

[23] Christine Wiedinmyer, SK Akagi, Robert J Yokelson, LK Emmons, JA Al-Saadi, JJ Orlando, and AJ Soja. The fire inventory from ncar (finn): A high resolution global model to estimate the emissions from open burning. *Geoscientific Model Development*, 4(3):625–641, 2011.

A Appendix

A.1 Chemical Fingerprinting: Procedure for Matching Daughter Mass Spectra with Chemical Compounds

Here we list the measurements we get when passing the jth sample through the UPLC instrument for $j = 1, \ldots, 71$.

1. Pairs of the first retention times and the parent mass $\left\{(t_{1,i}^{(j)}, m^{(j)}) : i = 1, \ldots, p_{all}\right\}$
2. Pairs of the second retention time and the mass spectrum $\left\{(t_{2,i}^{(j)}, \mathbf{v}_{\mathrm{MS},i}^{(j)}) : i = 1, \ldots, q_j\right\}$

Regarding 1, we get almost identical values for the first retention time and the parent mass in every 71 experiments. This allows us to set any of them as the representative values $\{(t_{1,i}, m_i) : i = 1, \ldots, p_{all}\}$. In the main article, we rather used the notation t_i for the first retention time. Since these two scalars are unique characteristic of the compounds, we can correspond one of each to the one of compounds in $\{\mathbf{x}_1, \ldots, \mathbf{x}_{p_{all}}\}$. It remains matching the compound with the mass spectrum $\mathbf{v}_{\mathrm{MS}}$, and we consider every MS measurements from all experiment. That is, we seek for a correspondence

$$f : \{\mathbf{x}_i = (t_{1,i}, m_i) : i = 1, \ldots, p_{\mathrm{all}}\} \longrightarrow \left\{\mathbf{v}_{\mathrm{MS},i}^{j} : i = 1, \ldots, q_j, j = 1, \ldots, p_{\mathrm{all}}\right\}.$$

To find this, we present the following heuristic criteria ordered by their importance.

1. Since the ionized compound travels the connecting column very fast, given the first retention time t_1 the second retention time t_2 must satisfy $t_1 < t_2 < t_1 + \delta$ for small $\delta > 0$.
2. Since MS shows the intensity of fragmented ions, there shall be no peak at any values greater than the parent mass m_i. This leads to censor the MS having 10% or greater of the total intensity measured is located after the parent mass.

If there is no MS that meets the above criteria, we increase the value of δ by degree of 0.01 until reaches $\delta = 0.1$. If multiple matches exist, we choose the MS closest to the representative spectrum as follows.

$$\mathbf{v}_{\mathrm{MS}} = \underset{\mathbf{v}^*:\mathrm{matched}}{\arg\min} \sum_{\mathbf{v}:\mathrm{matched}} \|\mathbf{v}^* - \mathbf{v}\|_2^2 .$$

To the end, we obtained 2,905 of compounds with full chemical fingerprints – a triplet of retention time, parent mass, and the mass spectrum – out of 3,319 compounds.

A.2 Traditional Statistical Distances between Probability Distributions

In this section, we present well-known statistical distances between two probability distributions.

A.2.1 Earth Mover's Distance (EMD)

The earth mover's distance (EMD), also known as the 1-Wasserstein distance, is a distance function between two probability distribution defined on a given metric space. Intuitively, the EMD metric is the minimum *cost* of moving one pile into the other, which is assumed to be the amount of pile that needs to be moved times the mean distance it has to be transported. Recall that the smoke sample can be viewed as a histogram over a set of compounds $\mathcal{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_p\}$. To compute the EMD between two smoke samples, we first need to define a metric on $\mathcal{X}$ which quantifies the cost moving one chemical compound to another compound. We call this metric – the distance between compounds in our case – as a ground distance. The chemical fingerprints of the compounds provides the possible choices of the ground distance. We can either use one of the retention time, the parent mass, and the mass spectrum (MS) because each of them conveys unique information of compounds.

The common definition of the p-Wasserstein between distributions $\mathbb{P}$ and $\mathbb{Q}$ on a metric space (M, d) is given by

$$W_p(\mathbb{P}, \mathbb{Q}) = \left(\inf_{\gamma \in \Gamma(\mathbb{P}, \mathbb{Q})} \mathbb{E}_{(x,y)\sim\gamma} d(x, y)^p\right)^{1/p},$$

where $\Gamma(\mathbb{P}, \mathbb{Q})$ denotes a collection of joint distribution of which marginal distributions are $\mathbb{P}$ and $\mathbb{Q}$. Given two smoke samples $\mathcal{D} = (\{(\mathbf{x}_1, w_{\mathbf{x}_1}), \ldots, (\mathbf{x}_p, w_{\mathbf{x}_p})\}, f)$ and $\mathcal{D}' = (\{(\mathbf{x}_1, w'_{\mathbf{x}_1}), \ldots, (\mathbf{x}_p, w'_{\mathbf{x}_p})\}, f')$, this optimization problem reduces to find a optimal $p \times p$ joint probability matrix whose row-wise and column-wise sum equals to the histogram representation of each of two smoke samples, respectively. That is,

$$\text{Minimize} \sum_{i=1}^{p}\sum_{j=1}^{p} h_{i,j} d(\mathbf{x}_i, \mathbf{x}_j) \quad \text{subject to} \quad \sum_{i=1}^{p} h_{i,j} = w_{\mathbf{x}_j}, \sum_{j=1}^{p} h_{i,j} = w'_{\mathbf{x}_i}, h_{i,j} \geq 0.$$

The $p \times p$ matrix $\mathbb{H} = (h_{i,j})_{i,=1}^{p}$ is called the optimal flow matrix such that each $h_{i,j}$ captures the optimal shift from compound $\mathbf{x}_i$ to compound $\mathbf{x}_j$. The EMD metric between $\mathcal{D}$ and $\mathcal{D}'$ is defined as the minimum of the given optimization problem.

A.2.2 Maximum Mean Discrepancy (MMD)

The maximum mean discrepancy (MMD) is a metric defined on a space of probability distributions. Let $\mathcal{H}$ be a reproducing kernel Hilbert space (RKHS) induced by a kernel function k or an implicit feature map ϕ where $k(x, y) = \langle \phi(x), \phi(y) \rangle_{\mathcal{H}}$. For a given probability distribution P, define a mean embedding $\mu_P \in \mathcal{H}$ satisfying that for all $f \in \mathcal{H}$,

$$\mathbb{E}_{X \sim P} f(X) = \langle f, \mu_p \rangle_{\mathcal{H}}.$$

The MMD of distributions P and Q computes the L^2-distance between their mean embeddings on $\mathcal{H}$.

$$\begin{aligned} \text{MMD}^2(P, Q) &= \|\mu_P - \mu_Q\|_{\mathcal{H}}^2 \\ &= \mathbb{E}_{X, X' \sim P} k(X, X') + \mathbb{E}_{Y, Y' \sim Q} k(Y, Y') - 2\mathbb{E}_{X \sim P, Y \sim Q} k(X, Y). \end{aligned}$$

[MMD] proposed a biased estimator of MMD. In particular, we are interested in computing the distance between the smoke samples. Let $\hat{F} = \{(\hat{f}_1, c_1), \ldots, (\hat{f}_m, c_m)\}$ and $\hat{G} = \{(\hat{g}_1, c_1), \ldots, (\hat{g}_m, c_m)\}$ be two smoke samples expressed as histograms over a common set of compounds. Then,

$$\begin{aligned} \widehat{\text{MMD}}^2(F, G) &= \sum_{i=1}^{m}\sum_{j=1}^{m} \hat{f}_i \hat{f}_j k(c_i, c_j) + \sum_{i=1}^{m}\sum_{j=1}^{m} \hat{g}_i \hat{g}_j k(c_i, c_j) - 2\sum_{i=1}^{m}\sum_{j=1}^{m} \hat{f}_i \hat{g}_j k(c_i, c_j) \\ &= (\hat{\mathbf{f}} - \hat{\mathbf{g}})^\top \mathbf{K} (\hat{\mathbf{f}} - \hat{\mathbf{g}}), \end{aligned}$$

where $\hat{\mathbf{f}} = (\hat{f}_1, \ldots, \hat{f}_m)^\top$, $\hat{\mathbf{g}} = (\hat{g}_1, \ldots, \hat{g}_m)^\top$, and $\mathbf{K}$ is an $m \times m$ kernel matrix of which (i, j)-th element is $k(c_i, c_j)$ for $i, j = 1, \ldots, m$.

Given a positive definite function w of two compounds, define a weight matrix $\mathbf{W} = (w(c_i, c_j))_{i,j=1}^{m}$. Let $\mathbf{D}$ be a diagonal matrix whose diagonal entry is a row-wise sum of $\mathbf{W}$, that is $\mathbf{D} = \text{Diag}(\{\sum_j w_{j,i}\})$. The transition matrix motivated by a random walk is given by $\mathbf{A} = \mathbf{D}^{-1}\mathbf{W}$. In the view of the spectral clustering, $\mathbf{A}$ is a normalized asymmetric graph Laplacian.

Remark 1. *Beside the nature of our data, another well-known graph Laplacian is a symmetric matrix* $\mathbf{D}^{-1/2}\mathbf{W}\mathbf{D}^{-1/2}$*. In this case, DDD is an exactly MMD with a kernel defined by*

$$k(u, v) = \frac{w(u, v)}{\sqrt{\sum_i w(u, c_i)}\sqrt{\sum_i w(v, c_i)}}.$$

A.2.3 Differentiating Smoke Samples via Traditional Optimal Transport

To calculate the EMD between distributions, we need to define a "ground distance" between chemical compounds as discussed in section A.2.1. The chemical characteristics of the compounds allow the possible choices of the ground distance. We can either use retention time or parent mass or mass spectrum (MS) because each of them conveys unique information of chemical compounds. Figure 13 depicts the MDS visualizations of fuel-specific smoke samples based on the pairwise EMD. Each panel shows the result based on EMD calculated using different ground distances, (a) Euclidean distance on the parent mass, (b) Euclidean distance on the retention time, and (c) the cosine distance on the daughter mass spectrum. Overall, EMDs calculated using MS representations of compounds more clearly distinguishes smoke samples belonging to three different classes compared to those using parent mass or retention time. The latter indicates that smoke samples tend to be compressed closer to each other, reflecting a smaller difference in discrepancy measurements. This can also be confirmed in Figure 2, for example, parent mass does not provide sufficient information to distinguish isomers. On the other hand, the high-dimensional MS representation of the compound prevents this degeneration. Since MS representation appears to result in more meaningful MDS visualization of the distance between smoke samples, we will use MS information in the remainder of the paper to represent compounds.

A.3 Multi-dimensional Scaling (MDS)

Multi-dimensional Scaling is one of the classic dimension reduction techniques for representing high-dimensional data in a lower-dimensional space. Given the pairwise distances between potentially high-dimensional instances of interest, MDS finds an optimal configuration of n points in $\mathbb{R}^k$ which approximately preserves the pairwise distances. Here k is the pre-chosen embedding dimension and is often set to $k = 2$ or $k = 3$ for visualization purposes. More precisely, let $\{l_{i,j} : i, j = 1, \ldots, n\}$ be a collection of pairwise distances, and suppose that there exist vectors $\mathbf{y}_1, \ldots, \mathbf{y}_n$ satisfying that $\|\mathbf{y}_i - \mathbf{y}_j\|_2 = l_{i,j}$. With the transitional invariance constraint $\sum_{i=1}^n \mathbf{y}_i = 0$, we can easily get

$$\mathbf{y}_i^\top \mathbf{y}_j = g_{i,j} := \frac{1}{2}\left(l_{\cdot,j}^2 + l_{i,\cdot}^2 - l_{\cdot,\cdot}^2 - l_{i,j}^2\right),$$

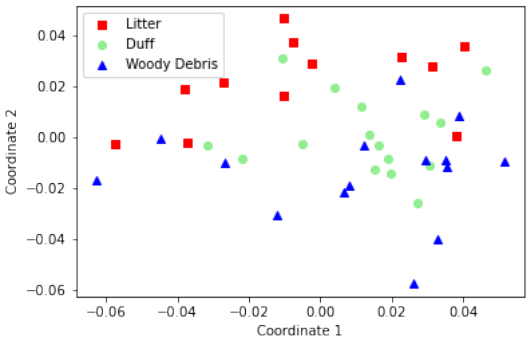
where $l_{\cdot,j}^2 = \sum_i l_{i,j}^2/n$, $l_{i,\cdot}^2 = \sum_j l_{i,j}^2/n$, and $l_{\cdot,\cdot}^2 = \sum_i \sum_j l_{i,j}^2/n^2$. This can be rewritten in a matrix form as $\mathbf{G} = \mathbf{Y}\mathbf{Y}^\top$ for a gram matrix $\mathbf{G} = (g_{i,j})_{i,j=1}^n$ and a embedding matrix $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_n]^\top$. Invoked by this, the solution of MDS is given by the best rank k approximation of $\mathbf{G}$. The MDS algorithm outputs $\mathbf{Y} = \mathbf{U}_k \Lambda_k$ where Λ_k is a $k \times k$ diagonal matrix which only collects the k largest eigenvalues of $\mathbf{G}$ and the columns of $\mathbf{U}_k$ are corresponding k eigenvectors of $\mathbf{G}$. The detailed algorithm can be found in, for example, [1].

A.4 Choice of Kernel Function to Define Similarity Between Compounds

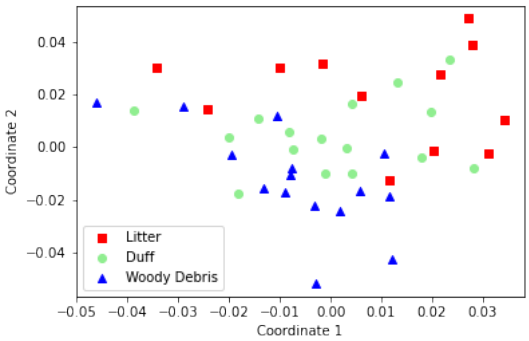
Our data set consists of 2,905 compounds each of which MS representation is a 10,000-dimension vector. As similarity function on the compounds, we choose the Gaussian kernel

$$w(x_i, x_j) = \exp\left(\frac{\|\mathbf{v}_{\mathbf{x}_i} - \mathbf{v}_{\mathbf{x}_i}\|_2^2}{2\sigma^2}\right).$$

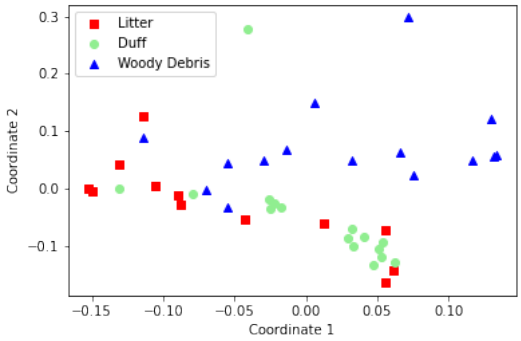
Here, we choose the median based bandwidth $\sigma = 0.52$ heuristically. In Figure 14a we show the top 10 eigenvalues of the row-wise normalized graph Laplacian. Each circle shown in Figure 14b corresponds to an individual compound.



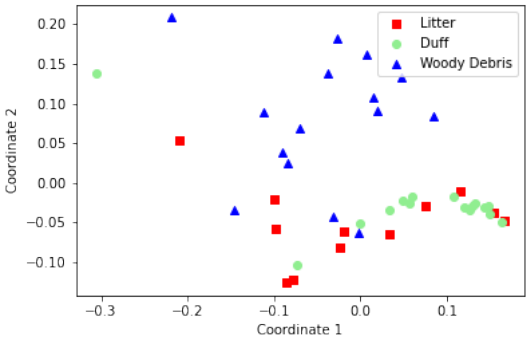
(a) Parent mass of matched compounds



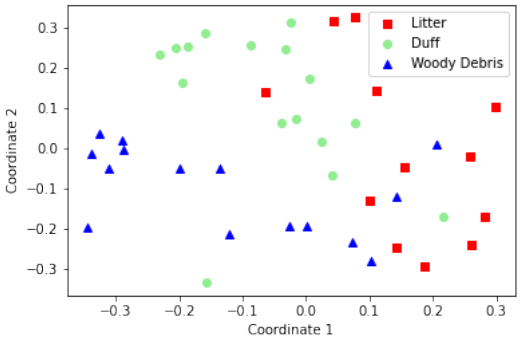
(b) Parent mass of all compounds



(c) Retention time of matched compounds

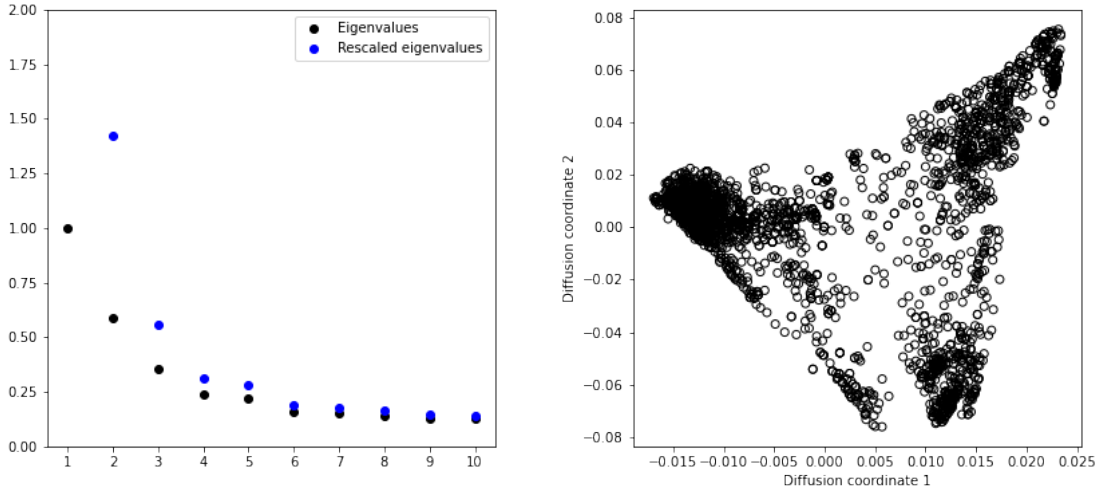


(d) Retention time of all compounds



(e) Mass spectrum of matched compounds

Figure 13: MDS representations of the EMD between smoke samples from three different fuels; Duff (yellow circle), Litter (red square), and Woody debris (blue triangle). To compute EMD for (a),(c), and (e), we only use the compounds which have matched MS2 information. Meanwhile, (b) and (d) are based on all compounds. The ground distances are defined as (a)(b) ℓ_1 distance on parent mass, (c)(d) ℓ_1 distance on retention time, and (e) the cosine distance on mass spectrum, respectively.



(a) (b)

Figure 14: We form the similarity matrix using the Gaussian kernel with a fixed bandwidth $\sigma = 0.52$ on the MS representations of compounds. Panel (a) is the scree plot of eigenvalues of the graph Laplacian. Panel (b) shows the 2-dimension visualization of the diffusion map of compounds.