NIST Technical Note
NIST TN 2297

Similarity Measures of Mass Spectra in Hilbert Spaces

Anthony J. Kearsley
Mathew J. Roberts

This publication is available free of charge from:
https://doi.org/10.6028/NIST.TN.2297

NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY
U.S. DEPARTMENT OF COMMERCE

NIST Technical Note
NIST TN 2297

Similarity Measures of Mass Spectra in Hilbert Spaces

Anthony J. Kearsley
Information Technology Laboratory
Applied and Computational Mathematics Division

Matthew J. Roberts
Institute for Defense Analyses
Cost Analysis and Research Division

This publication is available free of charge from:
https://doi.org/10.6028/NIST.TN.2297

July 2024

U.S. Department of Commerce
Gina M. Raimondo, Secretary

National Institute of Standards and Technology
Laurie E. Locascio, NIST Director and Under Secretary of Commerce for Standards and Technology

Certain equipment, instruments, software, or materials, commercial or non-commercial, are identified in this paper in order to specify the experimental procedure adequately. Such identification does not imply recommendation or endorsement of any product or service by NIST, nor does it imply that the materials or equipment identified are necessarily the best available for the purpose.

Publication History
Approved by the NIST Editorial Review Board on 2024-07-22

How to cite this NIST Technical Series Publication:
Kearsley AJ, Roberts MJ (2024) Similarity Measures of Mass Spectra in Hilbert Spaces. (National Institute of Standards and Technology, Gaithersburg, MD), NIST TN 2297. https://doi.org/10.6028/NIST.TN.2297

Author ORCID iDs
Anthony J. Kearsley: 0000-0002-9576-0621

Contact Information
Anthony.Kearsley@nist.gov

Abstract

Mass spectrometry (MS) is an analytic tool for measuring mass-to-charge ratios of molecules or molecular fragments. It is often used to identify and classify compounds. A mass spectrometer measures the mass of molecules and produces a signal proportional to the number detected fragments, thus, intensity can be interpreted as a function of mass-to-charge. One objective in mass spectrometry is distinguishing compounds using their measured mass spectra. In this technical report we describe the mathematical machinery required to compute a new similarity measure posed in a Hilbert space.

Keywords

Mass Spectrometry, Similarity Measure, Hilbert Space.

Table of Contents

List of Tables

List of Figures

Acknowledgments

The authors are grateful to NIST researchers, Drs. Amudhan Krishnaswamy Usha, Edward Sisco and Briana Capistran for invaluable input and a particular debt of gratitude is owed to Professor Arun S. Moorthy from Trent University for providing tireless and crucial guidance.

1. Section 1

Mass spectrometry (MS) is an analytic tool for measuring mass-to-charge ratios of molecules or molecular fragments. It is often used to identify and classify compounds. A mass spectrometer measures the mass of molecules and produces a signal proportional to the number detected fragments, thus, intensity can be interpreted as a function of mass-to-charge. Typically, one is interested in the peak structure in this signal [1, 2] and one objective in mass spectrometry is distinguishing compounds using their measured mass spectra.

A centroided mass spectrum, say m, is typically represented as ordered pairs of n positive real-valued scalars, $\{(x_1,y_1),(x_2,y_2)\dots(x_n,y_n)\}$ in which the first coordinate is the mass-to-charge ratio, and the second coordinate is the relative signal intensity. In practice, these relative intensities are most often determined by scaling the raw intensities $\tilde{y}_i$ by the maximum value of $\tilde{y}_i$. The point in the spectrum in which this maximum occurs is called the base peak. When scaled this way, the relative intensity of the base peak most often remains 1 from measurement to measurement. For that reason, we chose to normalize with respect to the 2-norm. More specifically,

$$y_i = \frac{\tilde{y}_i}{\sqrt{\sum_{j=1}^{n} \tilde{y}_j^2}}$$

While replicate measurements of the same compound or substance should be similar, they are never expected to be identical. Structurally similar compounds may produce very similar mass spectra, which can make distinguishing them extremely difficult.

A commonly employed, quantitative method for determining similarity between mass spectra is the *cosine similarity* measure in $\mathbb{R}^n$, also known as the dot product method [3],[4]. This is done by identifying the mass spectra of two compounds as vectors in $\mathbb{R}^n$, say v_m and v_s. The cosine similarity is computed by evaluating the cosine of the angle between them. There is no loss of generality to assume the vectors are of the same length, though in practice they may not be. The vectors v_m and v_s are determined by discretizing the x-axis, or mass-to-charge ratio values, into sub-intervals or bins I_k where $k = 1,2,\cdots,n$, and it is standard to form the bins I_k of uniform width.

The vector v_m is determined by the total relative signal intensity detected in each bin. More precisely, we define $m_i = \{(x,y) \in m \ : \ x \in I_i\}$ to be the subset of spectra in m to occur in bin I_i and we define the i^{th} component of v_m by

$$(v_m)_i = \sum_{(x,y)\in m_i} y. \tag{1}$$

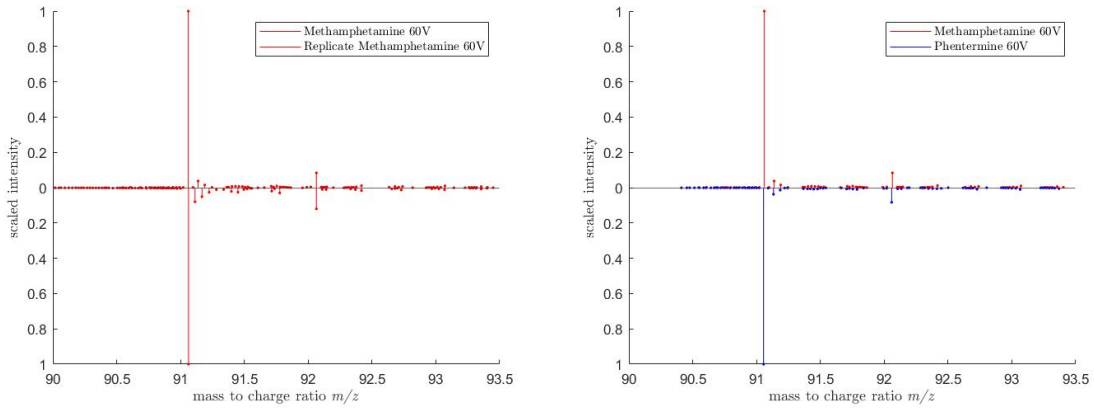

Fig. 1. Left: Replicate measurements of methamphetamine (a_1 and a_2). Due to significant noise, not every peak is representative of a molecule or molecular fragment. **Right:** Methamphetamine vs. phentermine (a_1 and b).

In the case that m_i is empty, $(v_m)_i = 0$. In the same way, the vector v_s is defined by the mass spectrum s and the cosine similarity $c(v_s, v_m)$ between v_s and v_m defined by

$$c(v_s, v_m) = \cos(\theta_{v_s,v_m}) = \frac{v_s \cdot v_m}{\|v_s\|_2 \|v_m\|_2}. \tag{2}$$

The cosine similarity measure is often employed because it is computationally inexpensive and it has been shown to be reliable in many applications [3]. However, in chemical identification, difficulties often arise when using this measure [5]. One such issue is miss-binning of spectra. It is possible for two associated peaks from two different mass spectra to be placed into different bins between replicate measurements. Many times this is due to poor calibration of the mass spectrometer, or poor selection of bin edges. Since cosine similarity is a measure of orthogonality, this means two replicate measurements of the same substance could be deemed to be dissimilar as they will have very low cosine similarity, [3],[4].

The main issue with the cosine similarity measure is its failure to distinguish certain pairs of structurally similar compounds. Measurement-to-measurement variations are investigated where structurally similar compounds produce similar mass spectra. Methamphetamine and phentermine are two structurally similar stimulants that are both controlled substances. Figure 1 shows a mass spectrum for phentermine in blue and two replicate mass spectra for methamphetamine in red. After binning, vectors $a_1, a_2 \in \mathbb{R}^n$ correspond to replicate mass spectra of methamphetamine and $b \in \mathbb{R}^n$ for phentermine. The cosine similarity between pairs a_1, a_2 and a_1, b are

$$c(a_1, a_2) = 0.9946\ ,\ c(a_1, b) = 0.9973, \tag{3}$$

which fails to distinguish methamphetamine from phentermine given only their mass spectra.

We recently explored incorporating measurement-to-measurement variability to improve the discriminability of near identical mass spectra of two structurally similar but unique compounds [6]. In this paper, we expand upon and generalize the preliminary work of consensus mass spectra [6],[7],[8],[9],[10], showing that one can think of the statistics acquired through replicate mass spectra as functions in a Hilbert space.

By utilizing replicate mass spectra, a collection of statistical parameters, say M, can be discovered related to the measurement m. Using these parameters, a mathematical model f_M can be defined for the underlying distribution from which m is sampled. The collection of statistical parameters, M, can be referred to as a *spectral statistic*. In general, the function model f_M will be an element of some appropriate Hilbert space, say H. Using these function models, one can define the generalized cosine similarity measure $\phi(M,S)$ in H between spectral statistics M and S by

$$\phi(M,S) = \cos(\theta_{f_M,f_S}) = \frac{\langle f_M, f_S \rangle_H}{\|f_M\|_H \|f_S\|_H}. \tag{4}$$

Here, θ_{f_M,f_S} is the angle between f_M and f_S in H, and will be compared to the standard cosine similarity measure c in $\mathbb{R}^n$ for pairs of structurally similar compounds.

Mass spectra generally fall into one of two types: High resolution and low resolution. High resolution mass spectra, are such that variation in the mass-to-charge ratios are recorded to high precision, unlike their low resolution counterpart. This gives rise to two distinct spectra statistics M: High dimensional consensus (HDC) spectra for high resolution mass spectra and discretized high dimensional consensus (dHDC) spectra for low resolution. These two types of spectral statistics lead to different Hilbert spaces in which the function model f_M exists. For this reason, it is natural to treat these two cases separately, but the objective for each case is the same: to generate a spectral statistic in which a function model can be defined, and use the cosine similarity between these function models to determine the similarity between spectral statistics.

2. High Resolution Mass Spectra

A high resolution mass spectrum is measured with an instrument that records mass-to-charge ratios and relative intensity values with high precision, allowing one to detect variance in both x and y values across replicate measurements. Therefore, it is natural to imagine each peak in a mass spectrum being sampled from a two dimensional probability distribution. In doing so, statistical parameters in two dimensions for each peak can be computed, and the collection of peak statistics forms an HDC spectrum. Each peak statistic P_i in an HDC spectrum M gives rise to a two dimensional probability distribution f_{P_i} in $L^2(\mathbb{R}^2)$. In turn, M can be identified as the linear combination $f_M = \sum_{i=1}^{n} a_i f_{P_i}$ of these

probability distributions in $H = L^2(\mathbb{R}^2)$. A similarity measure between HDC spectra M and S can then be defined by the cosine similarity between f_M and f_S in H.

2.1. Peak and Spectral Statistics

Let $m_1, m_2, \cdots, m_N$ be N replicate mass spectra for a single compound where each mass spectrum m_i is a collection of points in $\mathbb{R}^2$. Lengths of these spectra may vary depending on i, but we will assume they are all the same size. We denote p_1 as the peak of greatest relative intensity among all the replicate mass spectra. The peak from each spectrum m_i that is closest to peak p_1 in Euclidean distance can easily be determined, and from this one can form the set of peaks $S_1 = \{p_{1,1}, p_{1,2}, \cdots, p_{1,N}\}$, where $p_{1,j}$ is the peak from spectrum m_j. Here, $p_{1,i} = p_1$ since $p_1 \in m_i$. Thus S_1 consists of the replicate measurements of peak p_1 when grouped this way.

Once the set S_1 has been determined, these peaks can be removed from the spectra $m_1, m_2, \cdots, m_N$. Repeating this process with the newly defined spectra $m_1, m_2, \cdots, m_N$, the peak p_2 and the set $S_2 = \{p_{2,1}, p_{2,2}, \cdots, p_{2,N}\}$ are defined in the same fashion. Thus

$$S_j = \{p_{j,1}, p_{j,2}, \cdots, p_{j,N}\}, \tag{5}$$

where $p_{j,k}$ is the peak in m_k closest to peak p_j in Euclidean distance after removing peaks $p_{1,k}, p_{2,k}, \cdots p_{j-1,k}$ from the set m_k. The peaks $p_1, p_2, \cdots, p_n$ are referred to as *prominent* peaks.

Determining an appropriate or desirbale value of n, or equivalently, how many prominent peaks $p_1, p_2, \cdots, p_n$ should be considered when constructing sets $S_1, S_2, \cdots, S_n$ can be a challenge. In the analysis presented here, 20 prominent peaks are used, but in general the largest n peaks can be equal to the size of the smallest of the replicate spectra $m_1, m_2, \cdots, m_N$. One seeks to maximize n such that S_j, for any j, does not contain peaks that appear due exclusively to measurement noise. In essence, S_j is the set of replicate measurements of prominent peak p_j, thus the sample mean $\bar{p}_j = (\bar{p}_{x,j}, \bar{p}_{y,j})$ and the sample standard deviation $s_{p_j} = \left(s_{p_{x,j}}, s_{p_{y,j}}\right)$ of S_j yield important information about the location and variance of prominent peak p_j. The peak statistic P_j is then defined by the 4-tuple,

$$P_j = (\bar{p}_{x,j}, \bar{p}_{y,j}, s_{p_{x,j}}, s_{p_{y,j}}), \tag{6}$$

and the collection M of n peak statistics

$$M = \{P_1, P_2, \cdots, P_n\} \tag{7}$$

is the HDC mass spectrum of $m_1, m_2, \cdots, m_N$ using n prominent peaks [6].

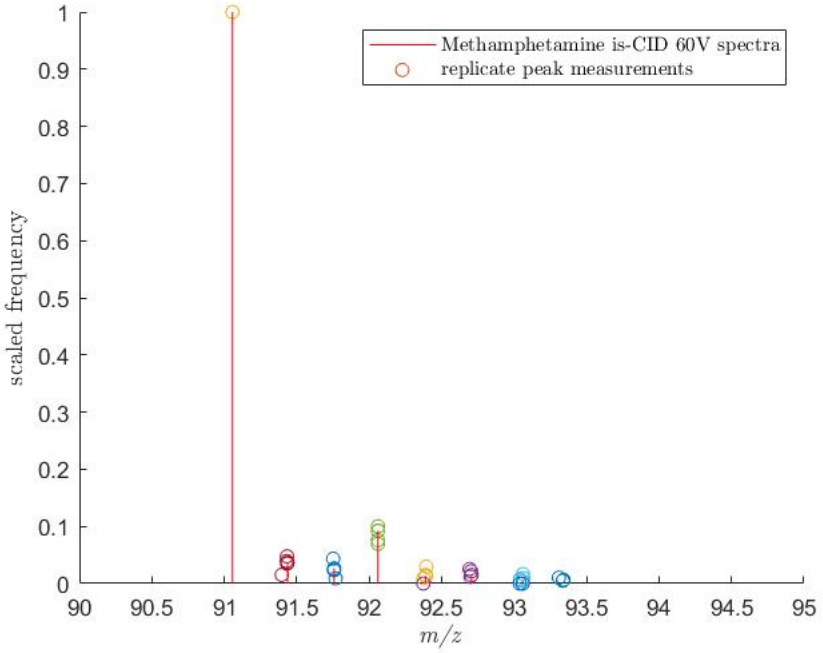

Fig. 2. Measured 60V spectra of three replicates for methamphetamine together with the sets S_j for the top 8 largest prominent peaks for methamphetamine for m/z 90 to m/z 95.

2.2. Peak Similarity with Peak Statistics

A peak statistic $P = (\bar{p}_x, \bar{p}_y, s_{p_x}, s_{p_y})$ can be identified with the 2D normal probability distribution

$$f_P(x,y) = \frac{1}{2\pi s_{p_x} s_{p_y}} e^{-\frac{1}{2}\left[\left(\frac{x-\bar{p}_x}{s_{p_x}}\right)^2 + \left(\frac{y-\bar{p}_y}{s_{p_y}}\right)^2\right]}. \tag{8}$$

Similarly, a peak statistic $Q = (\bar{q}_x, \bar{q}_y, s_{q_x}, s_{q_y})$ can be identified with

$$f_Q(x,y) = \frac{1}{2\pi s_{q_x} s_{q_y}} e^{-\frac{1}{2}\left[\left(\frac{x-\bar{q}_x}{s_{q_x}}\right)^2 + \left(\frac{y-\bar{q}_y}{s_{q_y}}\right)^2\right]}. \tag{9}$$

A natural way to determine the similarity between peak statistics P and Q is by computing the cosine similarity between f_p and f_q in the space $L^2(\mathbb{R}^2)$ [6]. The inner product of this space is given by

$$\langle f_P, f_Q \rangle_{L^2(\mathbb{R}^2)} = \iint_{\mathbb{R}^2} f_P(x,y) f_Q(x,y)\, dy\, dx. \tag{10}$$

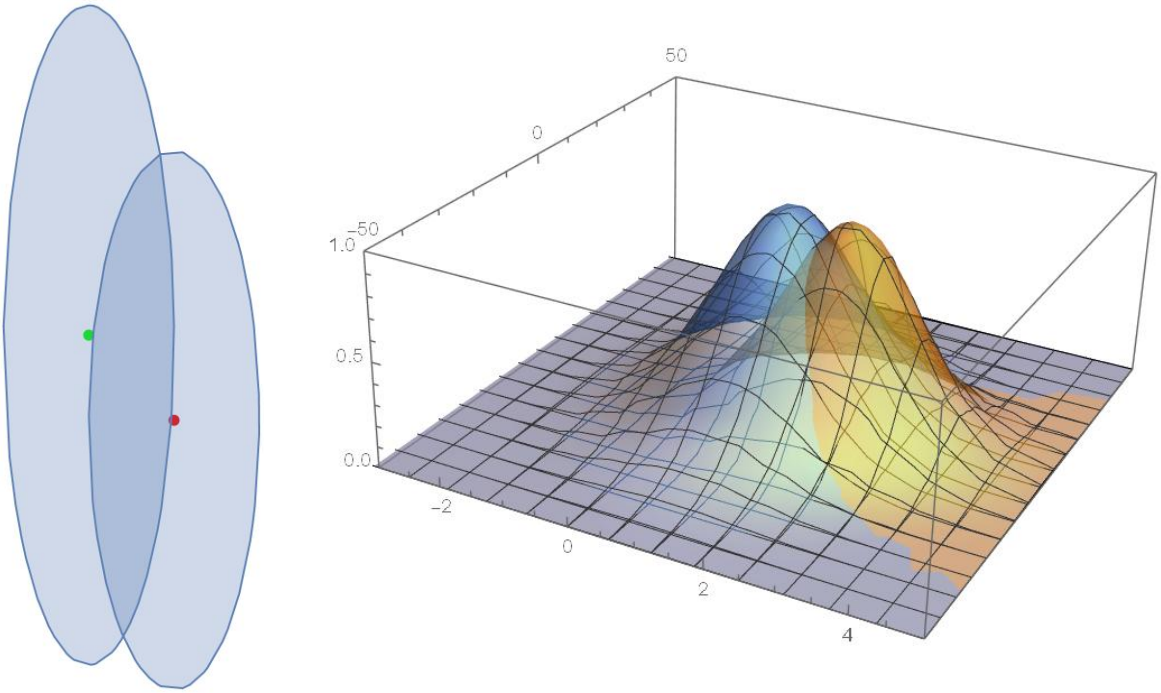

Fig. 3. Left: Peak statistics P and Q with mean and 1 standard deviation in x and y directions. **Red Point:** $(\bar{p}_x, \bar{p}_y)$, **Green Point:** $(\bar{q}_x, \bar{q}_y)$. **Right:** Plot of probability distributions $f_P(x,y)$ and $f_Q(x,y)$.

Therefore, the norm $\|f_P\|_{L^2(\mathbb{R}^2)}$ is given by

$$\|f_P\|_{L^2(\mathbb{R}^2)} = \sqrt{\langle f_P, f_P \rangle_{L^2(\mathbb{R}^2)}} = \sqrt{\iint\limits_{\mathbb{R}^2} f_P(x,y)^2 \, dy \, dx}. \tag{11}$$

The cosine of the angle between f_P and f_Q in $L^2(\mathbb{R}^2)$ can be written,

$$\cos(\theta_{f_P,f_Q}) = \frac{\langle f_P, f_Q \rangle_{L^2(\mathbb{R}^2)}}{\|f_P\|_{L^2(\mathbb{R}^2)} \|f_Q\|_{L^2(\mathbb{R}^2)}}$$

$$= \frac{\iint\limits_{\mathbb{R}^2} f_P(x,y) f_Q(x,y) \, dy \, dx}{\sqrt{\iint\limits_{\mathbb{R}^2} f_P(x,y)^2 \, dy \, dx} \sqrt{\iint\limits_{\mathbb{R}^2} f_Q(x,y)^2 \, dy \, dx}}.$$

Here, θ_{f_P,f_Q} is the angle between f_P and f_Q, and $\cos(\theta_{f_P,f_Q})$ is computed exactly by the formula

$$\cos(\theta_{f_P,f_Q}) = \sqrt{\frac{4 s_{p_x} s_{q_x} s_{p_y} s_{q_y}}{\left(s_{p_x}^2 + s_{q_x}^2\right)\left(s_{p_y}^2 + s_{q_y}^2\right)}} e^{-\frac{1}{2}\left[\left(\frac{\bar{p}_x - \bar{q}_x}{\sqrt{s_{p_x}^2 + s_{q_x}^2}}\right)^2 + \left(\frac{\bar{p}_y - \bar{q}_y}{\sqrt{s_{p_y}^2 + s_{q_y}^2}}\right)^2\right]}. \tag{12}$$

Separating this formula into x and y components, yields

$$\cos(\theta_{f_P,f_Q}) = \left(\sqrt{\frac{2 s_{p_x} s_{q_x}}{s_{p_x}^2 + s_{q_x}^2}} e^{-\frac{1}{2}\left(\frac{\bar{p}_x - \bar{q}_x}{\sqrt{s_{p_x}^2 + s_{q_x}^2}}\right)^2}\right)\left(\sqrt{\frac{2 s_{p_y} s_{q_y}}{s_{p_y}^2 + s_{q_y}^2}} e^{-\frac{1}{2}\left(\frac{\bar{p}_y - \bar{q}_y}{\sqrt{s_{p_y}^2 + s_{q_y}^2}}\right)^2}\right). \tag{13}$$

The similarity measure between two peak statistics P and Q, denoted by $\theta(P,Q)$, is given by the formula

$$\theta(P,Q) = \cos(\theta_{f_P,f_Q}), \tag{14}$$

takes on a value between 0 and 1. If f_P and f_Q are orthogonal in $L^2(\mathbb{R}^2)$, $\theta(P,Q) = 0$, $\theta(P,Q) \approx 0$ if f_P and f_Q nearly orthogonal and if f_P and f_Q are scalar multiples, then $\theta(P,Q) = 1$. Using the similarity measure θ, the similarity between function models f_M and f_S, can be defined as linear combinations in H, for HDC mass spectra M and S.

2.3. Similarity between HDC Spectra

Consider $M = \{P_1, P_2, \cdots, P_{n_M}\}$ and $S = \{Q_1, Q_2, \cdots, Q_{n_S}\}$, both HDC spectra. For each peak statistic $P_i = (\bar{p}_{x,i}, \bar{p}_{y,i}, s_{p_{x,i}}, s_{p_{y,i}})$, define $f_{P_i} \in L^2(\mathbb{R}^2)$ by

$$f_{P_i}(x,y) = \frac{1}{\sqrt{\pi s_{p_{x,i}} s_{p_{y,i}}}} e^{-\frac{1}{2}\left[\left(\frac{x-\bar{p}_{x,i}}{s_{p_{x,i}}}\right)^2 + \left(\frac{y-\bar{p}_{y,i}}{s_{p_{y,i}}}\right)^2\right]}. \tag{15}$$

The scalar-valued function f_{P_i} is similar to what was defined in Section 2.2 by equation (8), but differs in that it is normalized in $L^2(\mathbb{R}^2)$. Similarly for $Q_j = (\bar{q}_{x,j}, \bar{q}_{y,j}, s_{q_{x,j}}, s_{q_{y,j}})$ one can define

$$f_{Q_j}(x,y) = \frac{1}{\sqrt{\pi s_{q_{x,j}} s_{q_{y,j}}}} e^{-\frac{1}{2}\left[\left(\frac{x-\bar{q}_{x,j}}{s_{q_{x,j}}}\right)^2 + \left(\frac{y-\bar{q}_{y,j}}{s_{q_{y,j}}}\right)^2\right]}. \tag{16}$$

Since f_{P_i} and f_{Q_j} are normalized in $L^2(\mathbb{R}^2)$,

$$\langle f_{P_i}, f_{Q_j} \rangle_{L^2(\mathbb{R}^2)} = \frac{\langle f_{P_i}, f_{Q_j} \rangle_{L^2(\mathbb{R}^2)}}{\|f_{P_i}\|_{L^2(\mathbb{R}^2)} \|f_{Q_j}\|_{L^2(\mathbb{R}^2)}} = \theta(P_i, Q_j). \tag{17}$$

If f_M and f_S are defined as linear combinations of f_{P_i} and f_{Q_j} respectively, each weighted by the mean relative peak intensities $\bar{p}_{y,i}$ and $\bar{q}_{y,j}$, then

$$f_M = \sum_{i=1}^{n_M} \bar{p}_{y,i} f_{P_i}\ ,\ f_S = \sum_{j=1}^{n_S} \bar{q}_{y,j} f_{Q_j}. \tag{18}$$

One can define the similarity $\phi(M,S)$ between HDC spectra M and S to be the cosine similarity between f_M and f_S in $H = L^2(\mathbb{R}^2)$ yielding

$$\phi(M,S) = \cos(\theta_{f_M,f_S}) = \frac{\langle f_M, f_S \rangle_H}{\|f_M\|_H \|f_S\|_H}, \tag{19}$$

where, θ_{f_M,f_S} is the angle between f_M and f_S in H. To compute equation (19), it is advantageous to first compute $\langle f_M, f_S \rangle_H$,

$$\begin{aligned}\langle f_M, f_S \rangle_H &= \left\langle \sum_{i=1}^{n_M} \bar{p}_{y,i} f_{P_i}, \sum_{j=1}^{n_S} \bar{q}_{y,j} f_{Q_j} \right\rangle_H = \sum_{i=1}^{n_M} \sum_{j=1}^{n_S} \bar{p}_{y,i} \bar{q}_{y,j} \langle f_{P_i}, f_{Q_j} \rangle_H \\ &= \sum_{i=1}^{n_M} \sum_{j=1}^{n_S} \bar{p}_{y,i} \bar{q}_{y,j} \theta(P_i, Q_j).\end{aligned}$$

By a similar calculation,

$$\|f_M\|_H^2 = \langle f_M, f_M \rangle_H = \sum_{i=1}^{n_M} \sum_{j=1}^{n_M} \bar{p}_{y,i} \bar{p}_{y,j} \theta(P_i, P_j).$$

Putting these results together, one arrives at the formula

$$\phi(M,S) = \frac{\langle f_M, f_S \rangle_H}{\|f_M\|_H \|f_S\|_H} \tag{20}$$

$$= \frac{\sum_{i=1}^{n_M} \sum_{j=1}^{n_S} \bar{p}_{y,i} \bar{q}_{y,j} \theta(P_i, Q_j)}{\sqrt{\sum_{i=1}^{n_M} \sum_{j=1}^{n_M} \bar{p}_{y,i} \bar{p}_{y,j} \theta(P_i, P_j)} \sqrt{\sum_{i=1}^{n_S} \sum_{j=1}^{n_S} \bar{q}_{y,i} \bar{q}_{y,j} \theta(Q_i, Q_j)}}. \tag{21}$$

2.4. Peak Matching with HDC Spectra

Let $M = \{P_1, P_2, \cdots, P_{n_M}\}$ be an HDC mass spectrum such that the peak statistics P_i are well separated, that is

$$\delta = \left(\frac{\bar{p}_{x,i} - \bar{p}_{x,j}}{\sqrt{s^2_{p_{x,i}} + s^2_{p_{x,j}}}} \right)^2 + \left(\frac{\bar{p}_{y,i} - \bar{p}_{y,j}}{\sqrt{s^2_{p_{y,i}} + s^2_{p_{y,j}}}} \right)^2 \tag{22}$$

is larger than a prescribed tolerance, say D, when $i \neq j$. Therefore, for two distinct peak statistics P_i and P_j,

$$\theta(P_i, P_j) = \sqrt{\frac{4 s_{p_x} s_{q_x} s_{p_y} s_{q_y}}{\left(s^2_{p_x} + s^2_{q_x}\right)\left(s^2_{p_y} + s^2_{q_y}\right)}} e^{-\frac{1}{2}\left[\left(\frac{\bar{p}_x - \bar{q}_x}{\sqrt{s^2_{p_x} + s^2_{q_x}}} \right)^2 + \left(\frac{\bar{p}_y - \bar{q}_y}{\sqrt{s^2_{p_y} + s^2_{q_y}}} \right)^2 \right]}$$
$$\leq e^{-\frac{D}{2}} \approx 0.$$

Therefore,

$$\sum_{i=1}^{n_M} (\bar{p}_{y,i})^2 \theta(P_i, P_i) \approx \sum_{i=1}^{n_M} \sum_{j=1}^{n_S} \bar{p}_{y,i} \bar{q}_{y,j} \theta(P_i, Q_j),$$

and because $\theta(P_i, P_i) = 1$ for each i,

$$\|\bar{p}_y\|_2^2 = \sum_{i=1}^{n} \bar{p}^2_{y,i} = \sum_{i=1}^{n_M} \bar{p}^2_{y,i} \theta(P_i, P_i).$$

Here, $\bar{p}_y$ is the vector in $\mathbb{R}^{n_M}$ containing the average relative intensities for each prominent peak p_i. More precisely,

$$(\bar{p}_y)_i = \bar{p}_{y,i} \text{ , for } i = 1, 2, \cdots, n_M.$$

When the HDC mass spectra M and S are similar to each other, the peak statistics Q_j from S are similar to those P_i from M. From this it is fair to assume that for a given P_{i_1} from M, there is a Q_{j_1} from S such that $\theta(P_{i_1}, Q_{j_1}) \approx 1$, so

$$\|f_{P_{i_1}} - f_{Q_{j_1}}\|_H^2 = \langle f_{P_{i_1}}, f_{P_{i_1}} \rangle_H - 2\langle f_{P_{i_1}}, f_{Q_{j_1}} \rangle_H + \langle f_{Q_{j_1}}, f_{Q_{j_1}} \rangle_H \approx 0.$$

From this it can be concluded that for any $i \neq i_1$,

$$\begin{aligned}\theta(Q_{j_1}, P_i) = \langle f_{Q_{i_1}}, f_{P_i} \rangle_H &= \langle f_{Q_{i_1}} - f_{P_{i_1}}, f_{P_i} \rangle_H + \langle f_{P_{i_1}}, f_{P_i} \rangle_H \\ &\leq \|f_{Q_{i_1}} - f_{P_{i_1}}\|_H \|f_{P_i}\|_H + \left| \langle f_{P_{i_1}}, f_{P_i} \rangle_H \right| \approx 0.\end{aligned}$$

In this way, it follows that for a given peak P_{i_1} from M, there is a peak Q_{j_1} in S such that

$$\theta(P_{i_1}, Q_{j_1}) \approx 1 \text{ and } \theta(P_i, Q_{j_1}) \approx 0 \text{ for } i \neq i_1.$$

By matching peaks appropriately, the peak pairs (P_{i_k}, Q_{i_k}) can be determined and

$$\sum_{k=1}^{n} \bar{p}_{y,i_k} \bar{q}_{y,j_k} \theta(P_{i_k}, Q_{j_k}) \approx \sum_{i=1}^{n_M} \sum_{j=1}^{n_S} \bar{p}_{y,i} \bar{q}_{y,j} \theta(P_i, Q_j).$$

Here, n is less than or equal to $\min(n_M, n_S)$, leading to the approximation for $\theta(M, S)$:

$$\theta(M, S) \approx \frac{\sum_{k=1}^{n} \bar{p}_{y,i_k} \bar{q}_{y,j_k} \theta(P_{i_k}, Q_{j_k})}{\|p_y\|_2 \|q_y\|_2}. \tag{23}$$

In this case, $\bar{p}_y$ and $\bar{q}_y$ are determined by

$$(\bar{p}_y)_k = \bar{p}_{y,i_k} \ , \ (\bar{q}_y)_k = \bar{q}_{y,j_k}. \tag{24}$$

In most cases, equation (23) is a good approximation for $\theta(M, S)$ due to the high measurement precision of mass-to-charge ratios in mass spectrometry. Because of this, it is expected that each distinct prominent peak p_i will have small sample standard deviation $s_{p_{x,i}}$. Hence, when p_i is distinct from p_j, the value of δ in (22) will be large. When M and S are similar, the optimal pairing (P_{i_k}, Q_{j_k}) of peak statistics is made apparent and a greedy algorithm can be employed, beginning with the peak statistics of highest relative intensity [6].

3. Low Resolution Mass Spectra

In low resolution mass spectra, the mass-to-charge ratio values are reported with low precision, usually to one decimal place or an integer value. It is natural to identify low resolution mass spectra with vectors in $\mathbb{R}^n$ using the same discretization as the cosine similarity

method in $\mathbb{R}^n$. These resulting vectors are called *spectral* vectors or binned spectra. Utilizing replicate measurements of mass spectra, statistical parameters of a spectral vector, of length n, can be computed. Here, n is the number of subintervals used to discretize the spectral vector/mass spectra. It is reasonable to refer to the pair of n dimensional vectors that form these parameters, as a *discretized high dimensional consensus* (dHDC) mass spectrum of dimension n. A dHDC mass spectrum U of dimension n gives rise then to an underlying distribution f_U in the Hilbert space $[L^2(\mathbb{R})]^n = L^2(\mathbb{R}) \times L^2(\mathbb{R}) \times \cdots \times L^2(\mathbb{R})$. A similarity measure can then be defined between dHDC mass spectra U and V by using the cosine similarity between f_U and f_V in H.

Let $u_1, u_2, \cdots, u_N \in \mathbb{R}^n$ be N spectral vectors that are constructed from discretizing replicate low resolution mass spectra. Define the n dimensional dHDC mass spectrum $U \in \mathbb{R}^{n\times 2}$ by

$$U = (\bar{u}, s_u), \tag{25}$$

where, $\bar{u} \in \mathbb{R}^n$ and $s_u \in \mathbb{R}^n$ are the sample mean and sample standard deviation, respectively, of $u_1, u_2, \cdots, u_N$. The i^{th} row of U, $(\bar{u}_i, (s_u)_i)$, therefore contains the statistical parameters of the relative signal intensity observed over the i^{th} subinterval. The value of this intensity can be modeled by a one-dimensional probability distribution, for example,

$$g_U^i(x) = \frac{1}{s_{u_i}\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\bar{u}_i}{s_{u_i}}\right)^2}. \tag{26}$$

Normalizing this distribution in $L^2(\mathbb{R})$ yields,

$$f_U^i(x) = \frac{1}{\sqrt{s_{u_i}\sqrt{\pi}}} e^{-\frac{1}{2}\left(\frac{x-\bar{u}_i}{s_{u_i}}\right)^2}. \tag{27}$$

Similarly, if $V \in \mathbb{R}^{n\times 2}$ is the dHDC mass spectrum from spectral vectors $v_1, v_2, \cdots, v_N$, the i^{th} row of V, $(\bar{v}_i, (s_v)_i)$, gives rise to the scaled distribution

$$f_V^i(x) = \frac{1}{\sqrt{s_{v_i}\sqrt{\pi}}} e^{-\frac{1}{2}\left(\frac{x-\bar{v}_i}{s_{v_i}}\right)^2}. \tag{28}$$

The similarity between component i statistics $U_{i,*}$ and $V_{i,*}$ is determined by the cosine of the angle between f_U^i and f_V^i in $L^2(\mathbb{R})$ and is given by

$$\begin{aligned}\cos(\theta_{f_U^i, f_V^i}) &= \frac{\langle f_U^i, f_V^i\rangle_{L^2(\mathbb{R})}}{\|f_U^i\|_{L^2(\mathbb{R})}\|f_V^i\|_{L^2(\mathbb{R})}} = \langle f_U^i, f_V^i\rangle_{L^2(\mathbb{R})} = \int_{-\infty}^{\infty} f_U^i(x) f_V^i(x)\, dx \\ &= \sqrt{\frac{2 s_{u_i} s_{v_i}}{s_{u_i}^2 + s_{v_i}^2}}\, e^{-\frac{1}{2}\left(\frac{\bar{u}_i - \bar{v}_i}{\sqrt{s_{u_i}^2 + s_{v_i}^2}}\right)^2},\end{aligned}$$

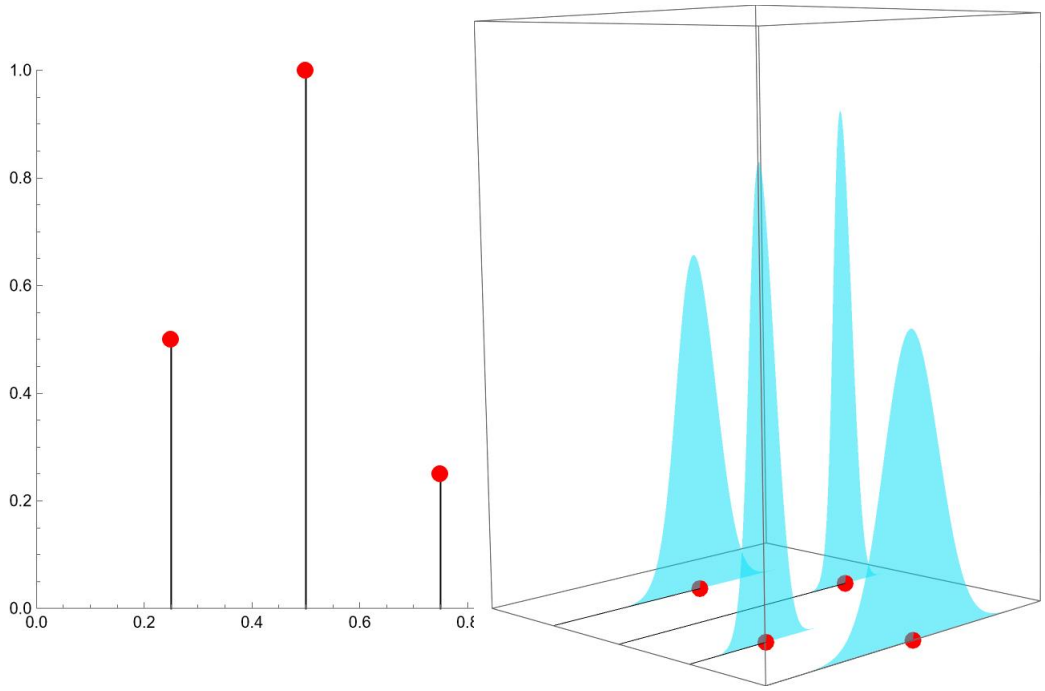

Fig. 4. Left: Artificial example of a discrete consensus mass spectrum with 4 peaks. **Right:** The same discrete consensus mass spectrum with scaled probability distributions f_U^i for each bin i.

where $\theta_{f_U^i, f_V^i}$ is the angle between f_U^i and f_V^i in $L^2(\mathbb{R})$.

One can identify n dimensional dHDC mass spectra in the Hilbert space $H = [L^2(\mathbb{R})]^n$, where for $f \in H$,

$$f = (f_1, f_2, \cdots, f_n) \ , \ f_i \in L^2(\mathbb{R}).$$

In this way, for $f, g \in H$, the inner product $\langle f, g \rangle_H$ is defined by

$$\langle f, g \rangle_H = \sum_{i=1}^{n} \langle f_i, g_i \rangle_{L^2(\mathbb{R})}. \tag{29}$$

For a dHDC mass spectrum U of dimension n, the function model, f_U, can be defined by

$$f_U = (\bar{u}_1 f_U^1, \bar{u}_2 f_U^2, \cdots, \bar{u}_n f_U^n). \tag{30}$$

Unlike HDC mass spectra, dHDC mass spectra represent discretized mass spectra, so each subinterval corresponds to a one-dimensional slice of the Cartesian plane. In this way, f_U represents the collection of scaled probability distributions for the relative intensity of mass spectra observed over each subinterval, with subintervals considered independently. Here it is useful to define the similarity measure $\psi(U,V)$ to be the cosine of the angle between f_U and f_V in H, thus,

$$\psi(U,V) = \cos(\theta_{f_U, f_V}) = \frac{\langle f_U, f_V \rangle_H}{\|f_U\|_H \|f_V\|_H}.$$

The numerator of this expression can be computed,

$$\langle f_U, f_V \rangle_H = \sum_{i=1}^{n} \langle \bar{u}_i f_U^i, \bar{v}_i f_V^i \rangle_{L^2(\mathbb{R})} = \sum_{i=1}^{n} \bar{u}_i \bar{v}_i \sqrt{\frac{2 s_{u_i} s_{v_i}}{s_{u_i}^2 + s_{v_i}^2}} e^{-\frac{1}{2}\left(\frac{\bar{u}_i - \bar{v}_i}{\sqrt{s_{u_i}^2 + s_{v_i}^2}}\right)^2},$$

and the denominator,

$$\|f_U\|_H^2 = \sum_{i=1}^{n} \langle \bar{u}_i f_U^i, \bar{u}_i f_U^i \rangle_{L^2(\mathbb{R})} = \sum_{i=1}^{n} \bar{u}_i^2 = \|\bar{u}\|_2^2.$$

Gathering these two things yields the formula,

$$\psi(U,V) = \frac{\sum_{i=1}^{n} \bar{u}_i \bar{v}_i \sqrt{\frac{2 s_{u_i} s_{v_i}}{s_{u_i}^2 + s_{v_i}^2}} e^{-\frac{1}{2}\left(\frac{\bar{u}_i - \bar{v}_i}{\sqrt{s_{u_i}^2 + s_{v_i}^2}}\right)^2}}{\|\bar{u}\|_2 \|\bar{v}\|_2}. \tag{31}$$

Equation (31) is similar to the cosine similarity $c(\bar{u}, \bar{v})$ between $\bar{u}$ and $\bar{v}$ given by equation (2). But with similarity measure ψ, each term $\bar{u}_i \bar{v}_i$ is weighted by the similarity between f_U^i and f_V^i in $L^2(\mathbb{R})$. In this way, the similarity measure ψ between dHDC mass spectra takes into account the statistical nature of mass spectra over each subinterval that is novel to the traditional methods that use consensus mass spectra [7],[8],[9],[10] [6].

Table 1. Table of 7 pairs of structurally similar compounds.

Pairs of Similar Compounds Considered	
1	Cotinine , Serotonin
2	Phenibut , MDA
3	MMDPPA , Methylone
4	5-methoxy MET , Norfentanyl
5	Cocaine , Scopolamine
6	HU-210 , Testosterone Isocaprionate
7	Methamphetamine, Phentermine

4. Numerical Results

Using the generalized cosine similarity method in the spaces $L^2(\mathbb{R}^2)$ and $[L^2(\mathbb{R})]^n$, 7 pairs of structurally similar compounds with similar spectra are considered. For each of these compounds, listed in Table 1, 30 measurements of mass spectra were collected at a fragmentation intensity of 30V using direct analysis in real time mass spectrometry (DART-MS), [11]. These mass spectra are high resolution, so both methods of generalized cosine similarity are applicable.

Using these replicate measurements, both HDC and dHDC mass spectra for each compound were generated. Replicate spectral statistics were computed using 15 of the 30 replicate spectra. For each pair of compounds, HDC spectra, quantities A_i and B_j are generated. By taking random subsets of 15 replicates from the collection of 30, 50 pairs of HDC spectra were produced in which each pair of HDC spectra were fashioned using disjoint sets of replicate mass spectra.

To assess the performance of the similarity measure ϕ, a comparison was made between the maximum values of $\phi(A_i, B_j)$ to the minimum values of $\phi(A_i, A_j)$ and $\phi(B_i, B_j)$. These values are recorded in Table 2. In this way, it can be determined if the similarity measure ϕ accurately distinguishes HDC spectra A_i and B_j from its replicates, or not. According to the min-max test developed by Moorthy and Sisco [12], the similarity measure ϕ successfully and accurately distinguishes every pair of structurally similar compounds.

Similarly, the same analysis can be performed, with similarity measure ψ when comparing dHDC spectra U_i and V_j. Here, U_i and V_j are replicate dHDC spectra constructed from each compound in the pair respectively. For each compound, 50 pairs of dHDC spectra were generated using random subsets of 15 replicate spectra from the total of 30 in the same way the HDC spectra were. The maximum similarity between $\psi(U_i, V_j)$ was then compared to the minimum similarity of $\psi(U_i, U_j)$ and $\psi(V_i, V_j)$. These values are displayed in Table 3.

Table 2. Table comparing the maximum of $\phi(A_i, B_j)$ to the minimum of $\phi(A_i, A_j)$ and $\phi(B_i, B_j)$. Here, A_i and B_j are replicate HDC spectra i and j coming from each compound respectively. Pairs of minimum and maximum values that fail the min-max test are displayed in bold.

Min and Max values ϕ	Maximum Value $\phi(A_i, B_j)$	Minimum Value $\phi(A_i, A_j)$ and $\phi(B_i, B_j)$
Cotinine and Serotonin	6.7575e-05	0.8380
Phenibut and MDA	0.0088	0.9287
MMDPA and Methylone	0.0010	0.6108
5-methoxy MET and Norfentanyl	0.9331	0.9491
Cocaine and Scopolamine	0.2435	0.6165
HU-210 and Testosterone	0.1538	0.6450
Methamphetamine and Phentermine	0.0150	0.3403

Table 3. Table comparing the maximum of $\psi(U_i, V_j)$ to the minimum of $\psi(U_i, U_j)$ and $\psi(V_i, V_j)$. Here, U_i and V_j are replicate dHDC spectra i and j coming from each compound respectively. Pairs of minimum and maximum values that fail the min-max test are displayed in bold.

Min and Max values ψ	Maximum Value $\psi(U_i, V_j)$	Minimum Value $\psi(U_i, U_j)$ and $\psi(V_i, V_j)$
Cotinine and Serotonin	8.1079e-04	0.8079
Phenibut and MDA	0.0096	0.8396
MMDPA and Methylone	0.1326	0.4053
5-methoxy MET and Norfentanyl	**0.9347**	**0.8714**
Cocaine and Scopolamine	0.2933	0.5693
HU-210 and Testosterone	0.2022	0.5935
Methamphetamine and Phentermine	0.2458	0.4187

Table 4. Table comparing the maximum of $c(a_i, b_j)$ to the minimum of $c(a_i, a_j)$ and $c(b_i, b_j)$. Pairs of minimum and maximum values that fail the min-max test are displayed in bold.

Min and Max values Cosine Similarity $\mathbb{R}^n$	Maximum Value $c(a_i, b_j)$	Minimum Value $c(a_i, a_j)$ and $c(b_i, b_j)$
Cotinine and Serotonin	0.3067	0.9414
Phenibut and MDA	0.4424	0.7438
MMDPA and Methylone	**0.4803**	**0.3404**
5-methoxy MET and Norfentanyl	**0.9700**	**0.9545**
Cocaine and Scopolamine	**0.9344**	**0.8339**
HU-210 and Testosterone	**0.9754**	**0.6962**
Methamphetamine and Phentermine	0.4956	0.7490

A comparison was then made between the performance of the traditional cosine similarity method using the dot product, and the new method proposed here. For a given pair of compounds from Table 1, the similarity of replicate measurements $a_1, a_2, \cdots, a_{30}$ of one compound was compared to replicate measurements $b_1, b_2, \cdots, b_{30}$ of the other compound. Here, the mass spectra a_i and b_j are identified as vectors in $\mathbb{R}^n$ using classical binning with a bin width of 0.1 and with the last bin from $899.9m/z$ to $900m/z$, thus $n = 9000$. The maximum cosine similarity score between any two a_i and b_j coming from the two sets of replicate measurements [12], was then considered. This similarity is determined by the formula

$$c(a_i, b_j) = \cos(\theta_{a_i,b_j}) = \frac{a_i \cdot b_j}{\|a_i\|_2 \|b_j\|_2}.$$

Here, θ_{a_i,b_j} is the angle between a_i and b_j in $\mathbb{R}^n$. To assess the performance of the cosine similarity measure, the maximum of $c(a_i, b_j)$ was compared to the minimum of $c(a_i, a_j)$ and $c(b_i, b_j)$. Results are displayed in Table 4. According to the min-max test, the cosine similarity measure frequently fails to accurately distinguish structurally similar compounds.

The similarity measures ϕ and ψ between HDC and dHDC spectra respectively outperform the classical cosine similarity method for discriminating structurally similar compounds. Computationally, the three methods are comparable. What is being leveraged, however, in the similarity measures ϕ and ψ, is the additional information coming from HDC and dHDC spectra.

5. Discussion and Conclusion

The use of replicate mass spectra to understand and utilize the statistical properties of measuring a mass spectrum is a novel idea in which a mass spectrum is replaced by the underlying probability distribution in which it is sampled from [6]. In this way, mass spectra

m and s are identified with the probability models f_M and f_S that incorporate the statistical behavior of measuring m and s, respectively. The concept of cosine similarity between mass spectra m and s can then be generalized to the cosine similarity between f_M and f_S in an appropriate Hilbert space.

The mass spectral statistics M and S generated from replicate measurements of m and s respectively determine the probability models f_M and f_S and two types of spectral statistics; HDC mass spectra for high resolution mass spectra and dHDC mass spectra for low resolution have been presented. In both cases, the position and variance of points in a mass spectrum are approximated leading to what appears to be an improved method of compound identification. To generate HDC mass spectra, points between replicate mass spectra were grouped according to Euclidean distance, where each group corresponds to replicate measurements of a particular peak in the spectrum. Using the statistical properties of each prominent peak, a regrouping could be performed to improve grouping of points as could various choices of distance in lieu of Euclidean distance. By utilizing other statistical parameters such as higher order moments, it may be possible to generate a probability model f_M that more accurately models the underlying distribution of mass spectral measurements for a mass spectrum.

Once the spectra statistics M and S have been collected, the similarity between M and S is determined by the cosine similarity between probability models f_M and f_S in an appropriate Hilbert space H. For HDC mass spectra and dHDC mass spectra, it appears natural to consider f_M and f_S as collections of normal distributions in the spaces $H = L^2(\mathbb{R}^2)$ and $H = [L^2(\mathbb{R})]^n$. Though these choices for f_M, f_S, and H yielded desirable results computationally, other choices for these objects may be more satisfactory depending on the spectral statistics M and S.

The use of replicate measurements to better understand the mass spectrum of a substance is an idea that is gaining popularity [6],[7],[8],[9],[10]. This method has been recently implemented [13] and employed to discriminate between terpene isomers [14]. The concept of identifying a mass spectrum with the probability distribution is a novel idea that has not been explored and may benefit from more advanced mathematical analysis.

References

[1] Wallace WE, Kearsley AJ, Guttman CM (2004) An operator-independent approach to mass spectral peak identification and integration. *Analytical Chemistry* 76(9):2446–2452.

[2] Kearsley AJ, Wallace WE, Bernal J, Guttman CM (2005) A numerical method for mass spectral data analysis. *Applied Mathematics Letters* 18(12):1412–1417.

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

[4] Wan KX, Vidavsky I, Gross ML (2002) Comparing similar spectra: from similarity index to spectral contrast angle. *Journal of the American Society for Mass Spectrometry* 13(1):85–88.

[5] Kearsley AJ, Moorthy AS (2021) Identifying fentanyl with mass spectral libraries. https://doi.org/10.26434/chemrxiv.14176952.v1. Available at https://doi.org/10.26434/chemrxiv.14176952.v1

[6] Roberts MJ, Moorthy AS, Sisco E, Kearsley AJ (2022) Incorporating measurement variability when comparing sets of high-resolution mass spectra. *Analytica Chimica Acta* 1230:340247.

[7] Nahnsen S, Bertsch A, Rahnenfuhrer J, Nordheim A, Kohlbacher O (2011) Probabilistic consensus scoring improves tandem mass spectrometry peptide identification. *Journal of Proteome Research* 10(8):3332–3343.

[8] Koo I, Zhang X, Kim S (2011) Wavelet-and fourier-transform-based spectrum similarity approaches to compound identification in gas chromatography/mass spectrometry. *Analytical Chemistry* 83(14):5631–5638.

[9] Place BJ (2021) Development of a data analysis tool to determine the measurement variability of consensus mass spectra. *Journal of the American Society for Mass Spectrometry* 32(3):707–715.

[10] Luo X, Bittremieux W, Griss J, Deutsch EW, Sachsenberg T, Levitsky LI, Ivanov MV, Bubis JA, Gabriels R, Webel H, et al. (2022) A comprehensive evaluation of consensus spectrum generation methods in proteomics. *Journal of Proteome Research* 21(6):1566–1574.

[11] Cody RB, Laramée JA, Nilles JM, Durst HD (2005) Direct analysis in real time (DART) mass spectrometry. *JEOL news* 40(1):8–12.

[12] Moorthy AS, Sisco E (2021) The min-max test: an objective method for discriminating mass spectra. *Analytical Chemistry* 93(39):13319–13325.

[13] Eveleth J, Moorthy AS, Kearsley AJ (2024) hdcms 0.1.27. Available at https://pypi.org/project/hdcms/.

[14] McGlynn DF, Andriamaharavo NR, Kearsley AJ (2024) Improved discrimination of mass spectral isomers using the high dimensional consensus mass spectral similarity algorithm. *Journal of Mass Spectrometry* :to appear.