Analytica Chimica Acta 1351 (2025) 343858

Contents lists available at ScienceDirect

Analytica Chimica Acta

journal homepage: www.elsevier.com/locate/aca

Novel probabilistic similarity scores for sets of replicate EI mass spectra

Amudhan Krishnaswamy-Usha [a,c], Briana A. Capistran [b], Anthony J. Kearsley [a,*]

[a] Applied and Computational Mathematics Division, National Institute of Standards and Technology, 100 Bureau Dr, Gaithersburg, 20899, MD, United States
[b] Materials Measurement Science Division, National Institute of Standards and Technology, 100 Bureau Dr, Gaithersburg, 20899, MD, United States
[c] Chakra Consulting Inc., 12004 Tregoning Pl, Clarksburg, 20871, MD, United States

ARTICLE INFO

Keywords:
Mass spectral similarity
Mass spectrometry
High dimensional consensus spectra
Kolmogorov–Smirnov scores

ABSTRACT

Background: Mass spectra are common signatures used to discriminate between compounds. This often involves the use of a *similarity score* to classify and distinguish between spectra of different compounds. To separate spectra of structurally similar compounds, multiple authors have explored the use of statistical and probabilistic methods applied to replicate mass spectra. In this paper, we explore the use of various averaged versions of the Kolmogorov–Smirnov and t-test statistics to compare peak intensities for sets of replicate mass spectra.

Results: Using replicate gas chromatography electron ionization (GC-EI-MS) mass spectra of 25 forensically relevant compounds, we compare the 'library match' and total classification accuracy of our novel probabilistic similarity scores versus the 'high dimensional consensus' (HDC) score and other similarity scores previously used in this context. We show that the use of the harmonic mean of the Kolmogorov–Smirnov test statistics obtained from peak intensities results in accuracies comparable to the HDC score.

Significance: Our results provide novel probabilistic similarity scores for replicate EI mass spectra, which outperform traditional scoring methods while at the same time minimizing the number of user-defined parameters and avoiding assumptions about the distributions of peak intensities.

1. Introduction

The identification of an unknown compound (an *analyte*) usually involves the comparison of the analyte's mass spectrum against a reference library of mass spectra. An analyst then uses various subjective or objective criteria to determine if the mass spectrum under consideration matches that of some compound in the library. Measurement of electron ionization mass spectra involves the inherently stochastic process of compound fragmentation. Even when controlling for other experimental factors, measurement variability induced by the stochastic nature of mass spectra complicates the task of objectively discriminating between structurally similar compounds, such as isomers. For instance, mass spectra of the isomers methamphetamine and phentermine are visually almost indistinguishable, even to a trained analyst (see Figs. 1 and 2).

Objective methods used to classify mass spectra typically involve the use of *similarity scores* or *match factors*, which assign a numerical value to a given pair of mass spectra (cf [1–3]). To tackle the problem of measurement variability, various authors have developed similarity scores or classification techniques which utilize replicate mass spectral measurements. See for instance the use of the 't-test' statistic in [4–7] and the construction of 'high-dimensional consensus' spectra in [8,9].

Building on these approaches, we construct a few replicate-based similarity scores based on common statistical tests, such as the Kolmogorov–Smirnov test and the t-test. In essence, these scores test the null hypothesis that two sets of spectra belong to the same underlying compound and assign a corresponding (pseudo) p-value. A distinguishing feature of these scores is that they are *parameter-free* and 'blind' - in other words, the scoring algorithms do not take additional information about the compounds into account. This avoids assumptions about the data, such as normality or heterogeneity of variance, which may not necessarily be true. Additionally, in the context of forensic applications, such scores potentially reduce the temptation an analyst might have to tweak parameters until they obtain a desired result.

We compare the efficacy of the scores introduced in this work to prior similarity scores, by evaluating them on a set of replicate EI spectra obtained from 25 compounds of forensic interest, including multiple sets of isomers (see Tables 1 and 2). The problem of assigning compound labels to mass spectra is an exercise in multi-class classification. There are a number of different notions of 'accuracy' one could define in this context. As we describe in Section 2.3, we focus on two

* Corresponding author.
E-mail address: Anthony.Kearsley@nist.gov (A.J. Kearsley).

https://doi.org/10.1016/j.aca.2025.343858
Received 19 January 2025; Received in revised form 21 February 2025; Accepted 22 February 2025
Available online 5 March 2025

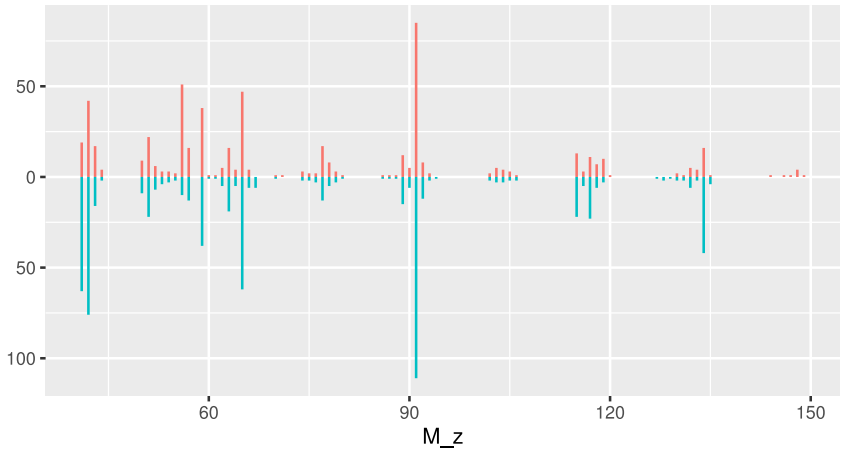

Fig. 1. Head to tail plot of mass spectra of methamphetamine (top) versus phentermine (bottom), depicted here with the base peak intensity at $m/z = 58$ removed. Intensities are normalized between 0 a.u. and 999 a.u.

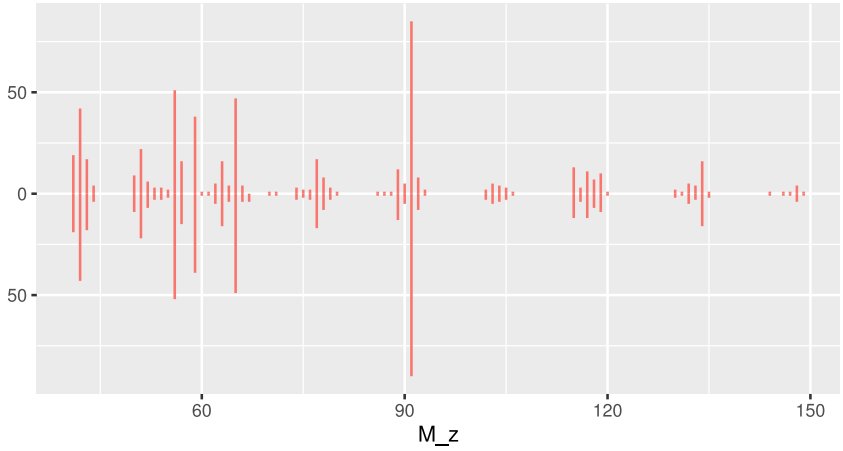

Fig. 2. Head to tail plot of mass spectra of two replicate spectra of methamphetamine, shown here with the base peak at $m/z = 58$ removed. Intensities are normalized between 0 a.u. and 999 a.u.

metrics which we believe are relevant from an end user's perspective — total classification accuracy and what we call the 'library match accuracy'. As we show in Section 3, our preliminary analysis shows that these scores outperform classical scores like cosine similarity and are comparable to other replicate spectra-based techniques.

Notation

We fix some notation for what follows in the rest of this work. We will use the letters x, y to denote mass spectra of compounds X, Y. Replicate spectra of compound X will be represented by superscripts $x^i, i = 1, \ldots, n$. Throughout this paper, we deal with 'low-resolution' mass spectra whose m/z values have been mapped to nominal integer values. Each mass spectrum x is therefore represented as a vector in some Euclidean space, with the tth component equal to the intensity value recorded for m/z value t. In what follows, we will use the letters s, t to denote m/z values and x_t to denote the intensity value of the spectrum at m/z value t. Unless otherwise specified, the intensity values are ℓ^2 normalized (i.e, $\sum_t (x_t)^2 = 1$).

A *similarity score* for us is a $[0, 1]$-valued function, which takes as input either pairs of individual spectra or sets of replicate spectra. All similarity scores will be denoted by Θ_A, with the subscript A indicating the type of score. $\Theta(x, y)$ is therefore the similarity score between spectra x and y. If a similarity score takes replicate spectra as input, this will be indicated by $\Theta((x^1, \ldots, x^n), (y^1, \ldots, y^m))$ or $\Theta((x^i), (y^j))$.

2. Methods

2.1. Computing similarity scores

The cosine similarity score is among a set of early similarity scores proposed for the comparison of mass spectra. Variants of the score (also known as 'match factors') are still in use today [1–3]. The cosine similarity score Θ_C between two spectra x and y is the normalized inner product between the vectors x, y in Euclidean space:

$$\Theta_C(x, y) = \frac{\sum_{t=1}^{M} x_t y_t}{\sqrt{\sum_t (x_t)^2}\sqrt{\sum_t (y_t)^2}}, \tag{1}$$

where M is some positive integer larger than the highest mass ion of x and y. Since our spectra are stored pre-normalized, this expression reduces to $\sum_t x_t y_t$.

Numerous authors have explored the use of statistical and probabilistic techniques to tackle measurement variability in mass spectra (cf. [4–6,8,10,11]). In [6] and subsequent work [4,5], a statistic based on Welch's unequal variance t-test is used to compare intensity values of sets of replicate spectra. Given two sets of intensity values $\{x_t^1, \ldots, x_t^n\}, \{y_t^1, \ldots, y_t^m\}$ at m/z value t, the t-test is used to test the null hypothesis that the means of the two sets are equal. In [6], two sets of spectra are deemed to be statistically equivalent if the null hypothesis is not rejected for any m/z value, at a given confidence level. In other

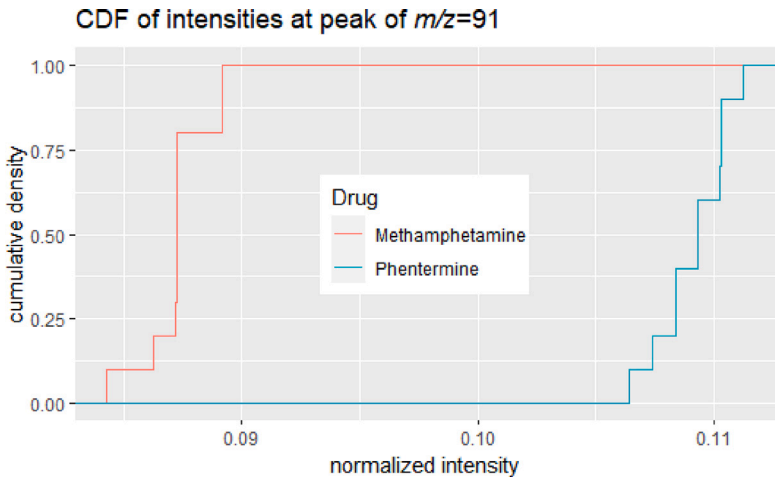

Fig. 3. Cumulative distribution functions of intensities measured at m/z value 91, from 10 replicate GC-EI-MS spectra of methamphetamine and phentermine. The maximum difference between these functions gives a KS test statistic of 1, with a corresponding p-value of 10^{-5}, reflecting the fact that the distributions of these intensities are different.

words, if p_t denotes the p-value obtained by an implementation of the t-test for m/z value t, this test concludes that spectra $\{x^i\}_i$ and $\{y^j\}_j$ represent different compounds if the minimum of these p_t falls below a pre-specified threshold. For the sake of comparison with the other similarity scores we introduce in this work, we construct a similarity score based on the classification technique outlined above.

Let $\{x^1, \ldots, x^n\}$ and $\{y^1, \ldots, y^m\}$ be replicate spectra of compounds X, Y. Let T denote the set of m/z values for which at least one of $\{x_t^1, \ldots, x_t^n, y_t^1, \ldots, y_t^m\}$ is non-zero. For each $t \in T$, apply Welch's t-test [12][1] to the sets $\{x_t^1, \ldots, x_t^n\}$, $\{y_t^1, \ldots, y_t^m\}$ to get a p-value p_t. We denote the 'score' obtained by taking the minimum of these p-values by $\Theta_{T_{min}}$, with

$$\Theta_{T_{min}}((x^i),(y^j)) = \min_{t \in T} p_t. \tag{2}$$

The score $\Theta_{T_{min}}$ is inspired by, but not strictly equivalent to, the technique described in [4]. The difference lies in the choice of m/z values to be compared - [4] and others pre-process the spectra to remove m/z values associated with known contaminants; whereas $\Theta_{T_{min}}$ utilizes the entire range of m/z values for the given sets of spectra.

We explore the use of other statistical tests to distinguish mass spectra of similar compounds. The two sample Kolmogorov–Smirnov (KS) test evaluates the null hypothesis that two sets of values $\{x_t^1, \ldots, x_t^n\}$ and $\{y_t^1, \ldots, y_t^m\}$ arose from the same underlying probability distribution, by comparing their empirical cumulative distribution functions (see Fig. 3). Similar to the computation of $\Theta_{T_{min}}$, we apply the KS test[1] to every set of $\{x_t^i\}, \{y_t^j\}$ where at least one of the spectra under consideration have non-zero intensity to get p-values p_t^{KS}. Multivariate analysis of relative ion abundances for sets of replicate spectra has shown that ion intensities are correlated across m/z values [14–16]. Therefore, the p-values obtained by applying statistical tests such as Kolmogorov–Smirnov (p_t^{KS}) and the t-test (p_t), are also correlated quantities. When faced with p-values from dependent statistical tests, a commonly used technique to combine them into a single p-value is to take their harmonic mean [17,18]. With this in mind, we define a 'KS based similarity score' $\Theta_{K_{hm}}$ to be the harmonic mean of the p-values p_t^{KS}.

$$\Theta_{K_{hm}}((x^i),(y^j)) = \frac{|T|}{\sum_{t \in T} \left(p_t^{KS}\right)^{-1}}, \tag{3}$$

where $|T|$ denotes the number of m/z values where the spectra have a non-zero intensity. Naturally, one could combine these p-values in various other ways — we explore the effects of taking the minimum, maximum, harmonic mean and arithmetic mean to the values p_t^{KS} or p_t to obtain analogous scores $\Theta_{K_{min}}, \Theta_{K_{max}}, \Theta_{K_{hm}}, \Theta_{K_{mean}}$ and $\Theta_{T_{min}}, \Theta_{T_{max}}, \Theta_{T_{hm}}, \Theta_{T_{mean}}$ respectively. We should note here that the resulting scores do not represent 'true' p-values - they are merely reasonable heuristics. Any claim about the statistical validity of these scores requires additional information and/or assumptions about the correlation between intensities for a given compound at distinct m/z values.

Alternative approaches to dealing with variability in measured mass spectra involve the construction of a 'consensus' mass spectrum, which represents some kind of average of the observed spectra [19]. In recent work involving one of the current authors [8], a notion of a 'high-dimensional consensus' (HDC) spectrum and associated HDC similarity score was introduced. Briefly, this involves viewing each mass spectrum as the realization of a mixture of 2-D Gaussian random variables, representing the mean position and variance of the m/z-intensity pairs (t, x_t). In other words, a HDC spectrum is a set of 4-tuples of the form $(\overline{t_1}, \overline{x_{t_1}}, \sigma_{t_1}, \sigma_{x_{t_1}})$, where $\overline{t_1}, \sigma_{t_1}$ and $\overline{x}_{t_1}, \sigma_{x_{t_1}}$ respectively represent the mean and variance of the m/z and intensity values of the largest peak in the spectra. Although a HDC spectrum (which we denote by $\overline{x}_{HD}$) with k peaks is from one perspective merely a vector of dimension $4k$, one may also think of a HDC spectrum as representing a set of Gaussian probability density functions, which live in the infinite dimensional function space $L^2(\mathbb{R}^2)$ (hence the 'high dimensional' adjective). The HDC score Θ_H between two sets of replicate spectra is obtained by computing a weighted average of the cosine similarity of these functions, viewed as elements in the Hilbert space $L^2(\mathbb{R}^2)$. In [9], a variant called the 'discrete HDC (dHDC) score' was introduced, for use in the case of low-resolution mass spectra. We note that the HDC score involves the choice of a parameter, namely the number of prominent peaks to be considered, while the dHDC score takes the entire mass spectrum into consideration. We will use Θ_{dH} to represent the dHDC score. For further details about the HDC and dHDC spectra and associated scoring algorithms, we refer the reader to [8,9,20]. (See also [21] for applications demonstrating the HDC score's utility in discriminating between isomers.)

2.2. Experimental data

We evaluate the scores described above on replicate EI spectra obtained from two sets of compounds of forensic interest, which we call 'Set A' and 'Set B'. Set A contains 16 compounds measured at constant concentrations while Set B consists of 9 isomers measured at varying concentration levels. The following describes the materials and instrumental analysis used for each set:

Unless otherwise noted, all chemicals were used as received.

[1] Implementations of which are widely available, for instance in R (R version 4.3.1 [13]).

Table 1
Seized drug compounds analyzed using GC-EI-MS, each 0.25 mg/mL in acetonitrile. ('Set A').

Compound	Formula
Acetyl fentanyl[a]	$C_{21}H_{26}N_2O$
Alprazolam	$C_{17}H_{13}ClN_4$
Amphetamine[a]	$C_9H_{13}N$
Benzyl fentanyl[a]	$C_{21}H_{26}N_2O$
Caffeine	$C_8H_{10}N_4O_2$
Cocaine	$C_{17}H_{21}NO_4$
Fentanyl	$C_{22}H_{28}N_2O$
Heroin	$C_{21}H_{23}NO_5$
Levamisole	$C_{11}H_{12}N_2S$
Methamphetamine[a]	$C_{10}H_{15}N$
Phentermine[a]	$C_{10}H_{15}N$
Phenylephrine[a]	$C_9H_{13}NO_2$
Quinine	$C_{20}H_{24}N_2O_2$
Tramadol[a]	$C_{16}H_{25}NO_2$
Trenbolone	$C_{18}H_{22}O_2$
Xylazine	$C_{12}H_{16}N_2S$

[a] Received as salt (HCl) form.

Table 2
Isomeric compounds, prepared at 0.33 mg/mL, 0.033 mg/mL, and 0.0033 mg/mL and analyzed by GC-EI-MS. ('Set B').

Compound	Formula
ortho-Fluorofentanyl[a] *meta*-Fluorofentanyl[a] *para*-Fluorofentanyl[a]	$C_{22}H_{27}FN_2O$
3,4-Methylenedioxy-α-methylaminohexaphenone[a] N-Methyl-N-propyl methylone	$C_{14}H_{19}NO_3$
MDMB-FUBINACA EMB-FUBINACA MDMB-FUB7AICA MDMB-4en-PINACA	$C_{22}H_{24}FN_3O_3$

[a] Received as salt (HCl) form.

Set A

Sixteen single-compound custom seized drug solutions (Table 1), each 0.25 mg/mL in acetonitrile, were purchased by Cayman Chemical (Ann Arbor, MI, USA). Aliquots of each solution were transferred to individual gas chromatograph (GC) vials, and ten replicates of each were analyzed. All samples were analyzed using an Agilent 8890 GC coupled to a 5977B mass selective detector (MSD) equipped with a 7693 autosampler (Agilent Technologies, Santa Clara, CA, USA). Analyses were performed using a DB-5 column (30 m length × 0.25 mm outer diameter × 0.25 μm inner diameter) and the following temperature program: initial temperature of 115 °C, hold for 1.5 min, ramp 10 °C/min to 135 °C, ramp 20 °C/min to 185 °C, ramp 14 °C/min to 290 °C, hold for 10 min. A 3 min solvent delay and 15:1 split ratio were used. The GC inlet was heated at 250 °C, and an injection volume of 1 μL was used. High-purity helium (99.999%) was used as the carrier gas for all analyses at a flow rate of 1.573 mL/min and pressure of 124.09 kPa (17.998 psi). The transfer line was set to 280 °C. For MS detection, the 70−eV electron ionization source was set to 250 °C, and the quadrupole was set to 150 °C. A gain of 1.00 arbitrary units (a.u.) was used, with a scan speed of 2.6 scans/s (N = 2). The scan range used was m/z 40–m/z 600 with a threshold of 150 counts. The extraction source tune (etune) was used for MS tuning.

Data acquisition was performed using MassHunter Workstation, GC/MS Data Acquisition software (version 10.2, Agilent Technologies). Retention times for all compounds were extracted at the apex of each chromatographic peak. Mass spectra (averaged across a given peak) were extracted using the National Institute of Standard and Technologies (NIST) Automated Mass Spectral Deconvolution and Identification System (AMDIS) (version 2.73, NIST) and the NIST Mass Spectral Search Program (version 2.3, NIST), equipped with the NIST/EPA/NIH 2017 EI Mass Spectral Library and SWGDRUG MS Library (version 3.9). Extracted spectra were automatically normalized between 0 a.u. and 999 a.u. in the Mass Spectral Search Program.

Set B

Nine seized drug compounds (Table 2), spanning three isomer classes, were all purchased in powder form from Cayman Chemical. Each powder was dissolved in 1 mL of methanol (≥ 99.9%, Sigma-Aldrich, St. Louis, MO, USA) to achieve a concentration of 1 mg/mL. The single-compound solutions were then serially diluted to achieve concentrations of 0.33 mg/mL, 0.033 mg/mL, and 0.0033 mg/mL, all in methanol. Ten replicate injections of each sample were analyzed via GC-EI-MS.

Compounds in Set B were analyzed using the same GC-EI-MS instrument used for Set A. These were analyzed a few months after the compounds in Set A. Any changes in experimental settings, such as the capillary column used, were necessitated by other experiments in progress at the time. Solutions were analyzed using a DB-35 column (30 m length × 0.25 mm outer diameter × 0.25 μm inner diameter) and the following temperature program: initial temperature of 115 °C, hold 1.5 min, ramp 10 °C/min to 135 °C, ramp 20 °C/min to 185 °C, ramp 14 °C/min to 290 °C, hold for 20 min. A 1.5 mL/min flow rate at 118.80 kPa (17.23 psi) was used for the carrier gas, which was high purity helium (99.999%). A 2 min solvent delay was employed, as well as a 1-μL injection volume (10:1 split ratio) for all injections. The GC inlet temperature was set at 250 °C. Ionization was performed through electron ionization (EI) (70 eV), and the source was set to 280 °C. The quadrupole was heated at 150 °C. The scan range of m/z 40–m/z 600 was used, at a scan speed of 5 scans/s (N = 1), as well as a threshold of 50 counts and gain of 1.00 arbitrary units (a.u.). MS tuning was performed using the extraction source tune (etune). We note here that etune is the tune typically used in forensic settings such as these, and that our comparison techniques remain valid as long as all compounds under consideration are analyzed using the same tune.

Chromatographic and mass spectral data was collected using MassHunter GC/MS Data Acquisition software. Normalized spectra were extracted using an in-house R script. Spectra were automatically normalized between 0 a.u. and 999 a.u.

2.3. Evaluating similarity scores

To evaluate the similarity scores listed above, we set a number n_L of 'library' replicates and a number n_A of 'analyte' replicates to be used. We then select n_L replicate spectra from each compound in our dataset and designate this set of mass spectra as our 'Library'. From the remaining spectra, we then sample n_A replicate spectra from each compound and designate this set as our 'Analytes'. We then compute the various similarity scores described above for the replicate EI spectra corresponding to each pair of compounds in the Library and Analyte sets. This computation is then repeated over a number of trials, corresponding to different sets of Library and Analyte spectra, with the final results taken as the average over the sets.

There are a number of related yet distinct ways of measuring the utility of a similarity score. We will focus on two metrics in our work - *total classification accuracy* and *library match accuracy*. Similarity scores are primarily used to make classification decisions based on a pre-selected threshold γ. That is, two mass spectra x and y (or sets of replicate spectra $(x^i), (y^j)$) are deemed to represent the same compound if and only if their score $\Theta(x, y)$ is greater than γ. The classification accuracy with threshold γ is then just the percentage of spectra correctly categorized in this manner (see Fig. 4 for an illustration). Ideally, one uses training data to select a threshold which maximizes this accuracy. We use our computed similarity scores to determine the optimal classification accuracy for every score type. An accuracy of 100% implies the scores are well separated - i.e., the minimal similarity score from

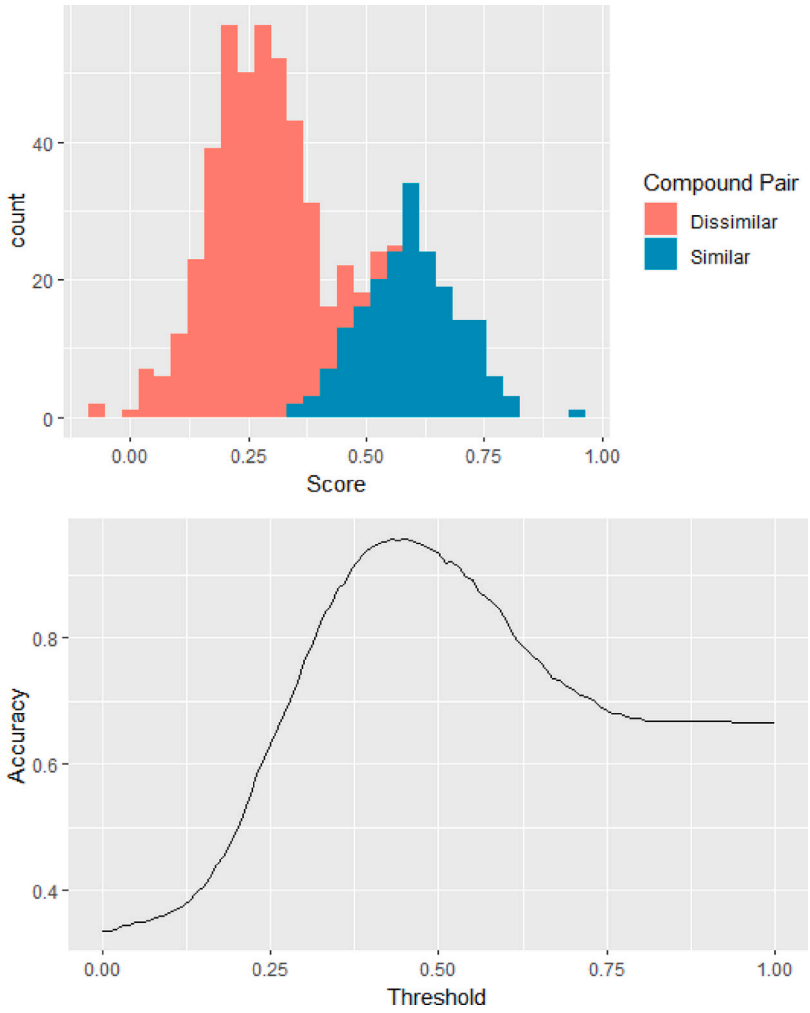

Fig. 4. Visualization of a hypothetical set of scores (top), along with a graph of accuracy as a function of the threshold (bottom). The maximum accuracy in this example is 95.6%, and occurs at a threshold of 0.46. Scores in blue depict scores between spectra from the same compound, while red scores are from dissimilar compounds. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

spectra of the same compound is greater than the maximum similarity score obtained from spectra belonging to different compounds. We call the difference between these two quantities the *inter-cluster separation*. Clearly, a positive inter-cluster separation implies 100% accuracy and vice versa.

In typical forensic applications, one compares a mass spectrum sample x against a library of spectra to generate a *hitlist* of likely matches [22]. A hitlist in this context is merely a list of mass spectra from the library, in descending order of their similarity score with x. We define the *library match accuracy* to be the fraction of analyte spectra for which the top match in the library hitlist belongs to the same compound. We note that despite some correlation with the total classification accuracy (for instance, classification accuracy of 100% implies a library match accuracy of 100%), they remain theoretically distinct concepts. We note that large libraries naturally increase the likelihood of false positives. On the other hand, one hopes that the use of multiple replicate spectra counteracts this trend. Further work is therefore needed to test the accuracy of these methods for larger libraries with a high number of replicates.

3. Results and discussion

Using 5 trials consisting of 6 replicate library spectra and 4 replicate analyte spectra from each compound, we evaluated the 11 similarity scores discussed above. The HDC score was computed using the 5 most prominent peaks for each compound. From these scores, we computed library match accuracy and the total classification accuracy, which are summarized in Fig. 5. Note that the use of 5 peaks is somewhat arbitrary, and merely serves to reduce computational time and provide a consistent baseline. Although some compounds under consideration, such as caffeine, produce several fragment ions, our analysis indicates that a partial spectrum of the top 5 peaks is sufficient to distinguish them. We hope to address this question in future work, and develop a noise model to eliminate this choice in the computation of a HDC spectrum. Apart from $\Theta_{T_{max}}$ and $\Theta_{K_{max}}$, the other replicate based scores presented in this work outperform traditional cosine similarity in terms of both metrics. We note that although the HDC and dHDC scores are optimal in terms of classification accuracy, they are outperformed by the t-test and Kolmogorov–Smirnov based variants $\Theta_{K_{hm}}, \Theta_{T_{mean}}$ and $\Theta_{K_{hm}}$ when it comes to library match accuracy. We currently do not have a satisfactory theoretical explanation of why this should be the case.

To disentangle the effects of concentration-dependent variation in the mass spectra, we repeat our analysis by restricting our dataset to the 16 compounds present in 'Set A', which were measured at a constant concentration of 0.25 mg/mL. As seen in Fig. 6, all score types apart from the *max* variants have 100% match accuracy. In other words, for each of these scores, the 'min–max distance' used in [23] is positive for each pair of compounds. In addition, four of these scores, namely $\Theta_{K_{mean}}, \Theta_{K_{hm}}, \Theta_{dH}$ and Θ_H also exhibit a 100% classification accuracy,

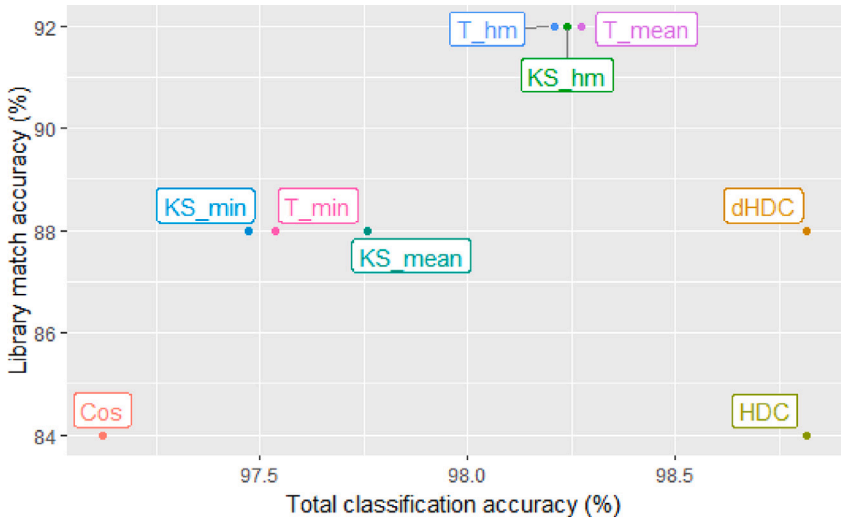

Fig. 5. Classification and library match accuracy for different score types evaluated on the data sets *A* and *B*. The K_{max} and T_{max} scores are not shown, as their accuracy is substantially lower.

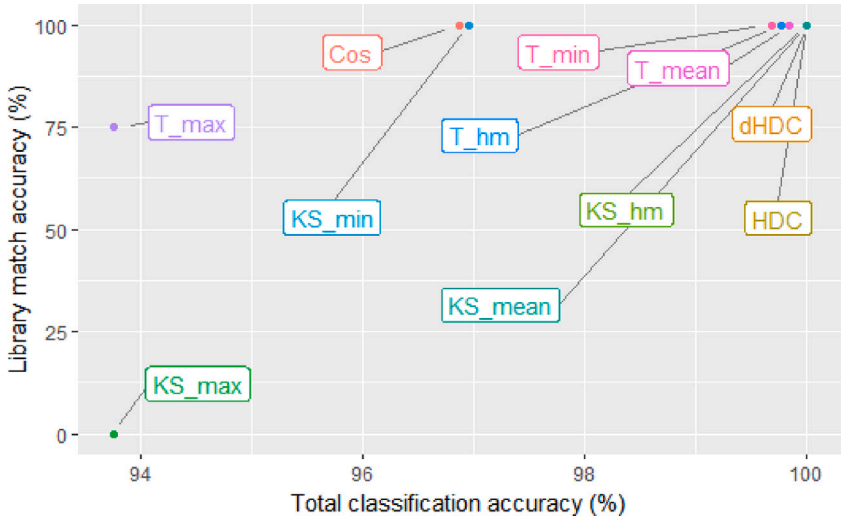

Fig. 6. Classification and library match accuracy for different score types evaluated on the constant concentration data set 'A', using 6 library replicates and 4 analyte replicates for each compound.

reflective of the fact that their *inter-cluster separation* is positive (Fig. 7). Classification using the $\Theta_{T_{max}}$ and $\Theta_{K_{max}}$ score is equivalent to the following procedure: apply a statistical test at every m/z value in the range of spectra x and y. If every resultant p-value is below a specified threshold, conclude that x and y belong to different compounds. Phrased this way, it should be clear that this increases the chance of Type I errors - i.e, the low accuracy of the *max* scores is due to a high rate of false positives.

Fig. 5 shows that replicate based similarity scores outperform cosine similarity. However, obtaining replicate spectra costs both time and money. One would therefore like to optimize the number of library and analyte replicates used to obtain a desired classification accuracy. The statistical power of the Kolmogorov–Smirnov test is very low for small sample sizes, but one might naively hope that the combined effect across multiple m/z values could counteract this. We tested the accuracy of the harmonic mean variant $\Theta_{K_{hm}}$ for library sizes ranging from 3 to 6 replicates against 1 to 4 analyte replicates (Fig. 8). For comparison, we note that the accuracy using cosine similarity is 97.1%. Our results depict clear increases in accuracy with larger analyte and library sizes, with $\Theta_{K_{hm}}$ outperforming cosine similarity for $n_L > 3$ and $n_A > 2$.

As shown in [9], the HDC score outperforms the dHDC score when applied to high resolution mass spectra. To a certain extent, this is because the process of obtaining low-resolution spectra involves a loss of information. In future work, we hope to investigate variants of the Θ_K and Θ_T scores suitable for use with high resolution spectra. Ideally, the additional information present in such mass spectra will improve classification and library match accuracy, even when faced with samples containing mixed concentrations. Alternatively, to deal with the problem of concentration-dependent variance, one could build a library of mass spectra measured at a range of concentrations. Given an analyte measured at some concentration within this range, one would then interpolate from the library to obtain appropriate spectra at the same concentration level. In ongoing work, we also seek to construct variants of the HDC score which require only one replicate EI spectrum as an analyte compound. Finally, we note that the similarity scores presented in this work are all one-dimensional in nature. In other words, we have not incorporated other information from the compounds beside their spectra, such as their chromatographs/retention indices. The addition of such extra dimensions is likely to increase accuracy, and is a subject we leave for future work.

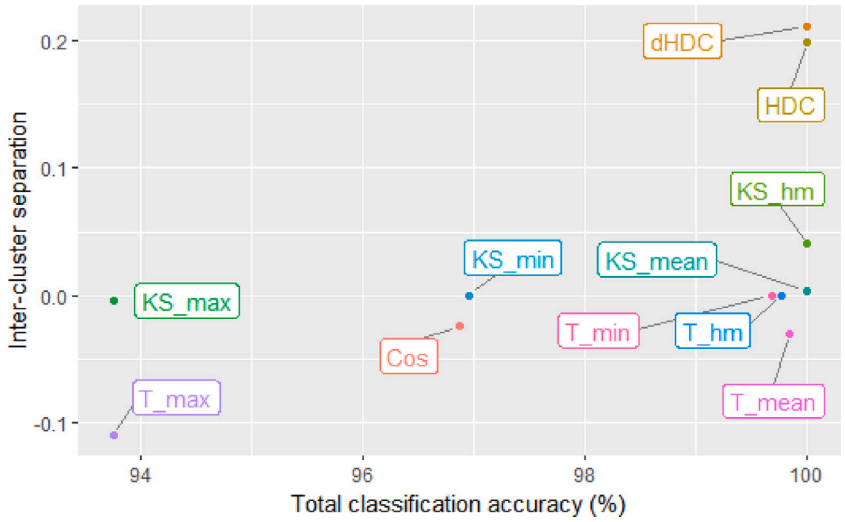

Fig. 7. Classification accuracy and inter-cluster separation for different score types evaluated on the constant concentration data set 'A'.

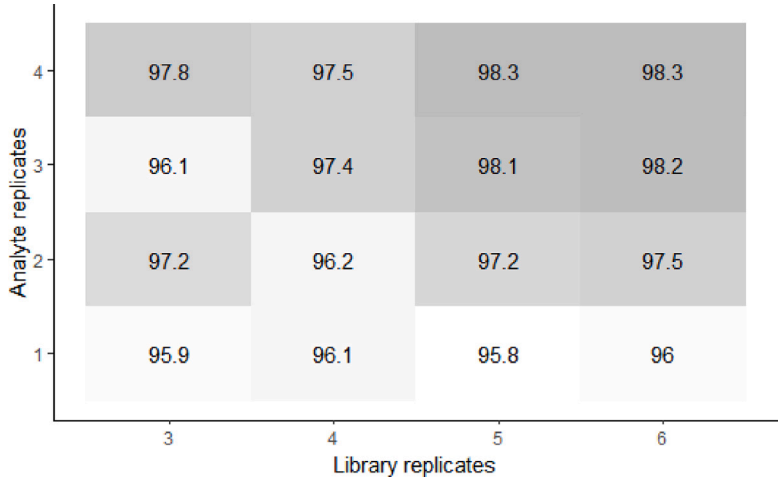

Fig. 8. Classification accuracy for the $\Theta_{KS_{hm}}$ score, as a function of the number of library and analyte replicates used.

4. Conclusion

We have introduced a few novel probabilistic similarity scores for sets of replicate EI spectra. These scores utilize the Kolmogorov–Smirnov and t-test statistics to compare the distribution of intensities at each m/z value in the range of the two sets of spectra. The resulting p-values obtained by applying these tests are then averaged by computing either the arithmetic mean, minimum, maximum or harmonic mean. By applying our scores to replicate spectra obtained from a set of 25 compounds, we conclude that the harmonic mean version of these scores is optimal, and has accuracy comparable to other recently introduced scores [8,9] based on 'high-dimensional consensus' spectra.

CRediT authorship contribution statement

Amudhan Krishnaswamy-Usha: Writing – review & editing, Writing – original draft, Software, Methodology, Investigation, Conceptualization. **Briana A. Capistran:** Writing – review & editing, Writing – original draft, Investigation, Data curation. **Anthony J. Kearsley:** Writing – review & editing, Writing – original draft, Supervision.

Disclaimer

Official contribution of the National Institute of Standards and Technology (NIST); not subject to copyright in the United States. Certain commercial products are identified in order to adequately specify the procedure; this does not imply endorsement or recommendation by NIST, nor does it imply that such products are necessarily the best available for the purpose.

Funding

K-U received funding through NIST, United States Task Order #: 1333ND
23FNB770067.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

References

[1] F. McLafferty, R. Hertel, R. Villwock, Probability based matching of mass spectra. Rapid identification of specific compounds in mixtures, Org. Mass Spectrom. 9 (1974) 690–702.

[2] S.E. Stein, D.R. Scott, Optimization and testing of mass spectral library search algorithms for compound identification, J. Am. Soc. Mass Spectrom. 5 (1994) 859–866.

[3] A.S. Moorthy, A.J. Kearsley, Pattern similarity measures applied to mass spectra, in: Progress in Industrial Mathematics: Success Stories: The Industry and the Academia Points of View, Springer, 2021, pp. 43–53.

[4] M.A.B. Willard, R.W. Smith, V.L. McGuffin, Statistical approach to establish equivalence of unabbreviated mass spectra, Rapid Commun. Mass Spectrom. 28 (2014) 83–95.

[5] A.M. Sacha, I.C. Willis, V.L. McGuffin, R. Waddell Smith, Identifying reliable ions for the statistical differentiation of structurally similar fentanyl analogs, J. Forensic Sci. 68 (2023) 1527–1541.

[6] M.A.B. Willard, V.L. McGuffin, R.W. Smith, Statistical comparison of mass spectra for identification of amphetamine-type stimulants, Forensic Sci. Int. 270 (2017) 111–120.

[7] E.L. Stuhmer, V.L. McGuffin, R.W. Smith, Discrimination of seized drug positional isomers based on statistical comparison of electron-ionization mass spectra, Forensic Chem. 20 (2020) 100261.

[8] M.J. Roberts, A.S. Moorthy, E. Sisco, A.J. Kearsley, Incorporating measurement variability when comparing sets of high-resolution mass spectra, Anal. Chim. Acta 1230 (2022) 340247.

[9] A.J. Kearsley, M. Roberts, Similarity Measures of Mass Spectra in Hilbert Spaces, Technical Note (NIST TN), National Institute of Standards and Technology, 2024.

[10] S. Nahnsen, A. Bertsch, J. Rahnenfuhrer, A. Nordheim, O. Kohlbacher, Probabilistic consensus scoring improves tandem mass spectrometry peptide identification, J. Proteome Res. 10 (2011) 3332–3343.

[11] R. Jain, M. Wagner, Kolmogorov-Smirnov scores and intrinsic mass tolerances for peptide mass fingerprinting, J. Proteome Res. 9 (2010) 737–742.

[12] B.L. Welch, The generalization of 'students's' problem when several different population variances are involved, Biometrika 34 (1947) 28–35.

[13] R. Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, 2023, URL: https://www.R-project.org/.

[14] J.T. Davidson, G.P. Jackson, The differentiation of 2 5-dimethoxy-n-(n-methoxybenzyl) phenethylamine (nbome) isomers using gc retention indices and multivariate analysis of ion abundances in electron ionization mass spectra, Forensic Chem. 14 (2019) 100160.

[15] S.A. Mehnert, J.T. Davidson, A. Adeoye, B.D. Lowe, E.A. Ruiz, J.R. King, G.P. Jackson, Expert algorithm for substance identification using mass spectrometry: application to the identification of cocaine on different instruments using binary classification models, J. Am. Soc. Mass Spectrom. 34 (2023) 1235–1247.

[16] G.P. Jackson, S.A. Mehnert, J.T. Davidson, B.D. Lowe, E.A. Ruiz, J.R. King, Expert algorithm for substance identification using mass spectrometry: Statistical foundations in unimolecular reaction rate theory, J. Am. Soc. Mass Spectrom. 34 (2023) 1248–1262.

[17] I.J. Good, Significance tests in parallel and in series, J. Amer. Statist. Assoc. 53 (1958) 799–813.

[18] D.J. Wilson, The harmonic mean p-value for combining dependent tests, Proc. Natl. Acad. Sci. 116 (2019) 1195–1200.

[19] X. Luo, W. Bittremieux, J. Griss, E.W. Deutsch, T. Sachsenberg, L.I. Levitsky, M.V. Ivanov, J.A. Bubis, R. Gabriels, H. Webel, et al., A comprehensive evaluation of consensus spectrum generation methods in proteomics, J. Proteome Res. 21 (2022) 1566–1574.

[20] A.J. Kearsley, HDCMS: A Package for Computing High-Dimensional Consensus Mass Spectral Similarity Scores, Technical Note (NIST TN), National Institute of Standards and Technology, 2025.

[21] D.F. McGlynn, N. Rabe Andriamaharavo, A.J. Kearsley, Improved discrimination of mass spectral isomers using the high-dimensional consensus mass spectral similarity algorithm, J. Mass Spectrom. 59 (2024) e5084.

[22] W.E. Wallace, A.S. Moorthy, Nist mass spectrometry data center standard reference libraries and software tools: Application to seized drug analysis, J. Forensic Sci. 68 (2023) 1484–1493.

[23] A.S. Moorthy, E. Sisco, The min–max test: an objective method for discriminating mass spectra, Anal. Chem. 93 (2021) 13319–13325.