The min-max test: an objective method for discriminating mass spectra

Arun S. Moorthy[†] & Edward Sisco

National Institute of Standards and Technology, Gaithersburg, MD, USA

Abstract: Deciding whether the mass spectra of seized drug evidence and a reference standard are measurements of two different compounds is a central challenge in forensic chemistry. Normally, an analyst will compute a mass spectral similarity score between spectra from the sample and reference and make a judgment using both the score and their visual interpretation of the spectra. This approach is inherently subjective and not ideal when rapid assessment of several samples is necessary. Making decisions using only the score and a threshold value greatly improves analysis throughput and removes analyst-to-analyst subjectivity, but selecting an appropriate threshold is itself a non-trivial task. In this manuscript, we describe and evaluate the min-max test – a simple and objective method for classifying mass spectra that leverages replicate measurements from each sample to remove analyst subjectivity. We demonstrate that the min-max test has an intuitive interpretation for decision-making, and its performance exceeds thresholding with similarity scores even when the best performing threshold for a fixed dataset is prescribed. Determining whether the underlying framework of the min-max test can incorporate retention indices for objectively deciding whether spectra are measurements of the same compound is on-going work.

Keywords: binary classification; electron ionization mass spectrometry; forensics; mass spectral interpretation; mass spectral libraries; min-max test; seized drug analysis.

1. Introduction

"Is it a fentanyl?" Drug analysts are routinely tasked with answering questions of this nature when presented with seized drug evidence. The most commonly employed analytical technique towards confirmatory tasks in seized drug analysis is gas chromatography mass spectrometry[1]; gas chromatography is used to physically and temporally separate the case sample into constituent components, and electron ionization (EI) mass spectrometry is used to propose the molecular structure (identity) of each component.

A mass spectrum can be thought of as a *roughly reproducible*[a] representation of a compound's structural information. In some cases, a mass spectrum includes near complete structural information, allowing it to be directly interpreted and the compound to be uniquely identified (see

[†] To whom correspondence should be addressed: arun.moorthy@nist.gov

[a] We introduce the phrase *roughly reproducible* to indicate that we expect replicate measurements to be self-similar, but we never expect replicate measurements to be identical.

introductory examples in references[2–4] among other textbooks). In most cases, however, mass spectra contain incomplete and non-unique structural information that render interpretation impractical without comparison to reference spectra. And while many mass spectral libraries[5,6] and interactive software tools[7,8] are available to assist in the interpretation process, the burden of decision-making still lies with the analyst.[9] In an application area like seized drug analysis, where decisions must be made rapidly and analyst subjectivity can be of significant consequence, having an explainable numerical approach for deciding whether mass spectra are measurements of different compounds is an obvious need.

We use the term *similarity score* to represent any numerical index that estimates the similarity between a pair of mass spectra. Although inconsistent in terminology and notation, several similarity scores have been explored in the literature.[10–17] The most well-known in mass spectrometry is the *dot product*,[11] or more commonly known as the *cosine similarity* in other pattern recognition applications. The cosine similarity between mass spectra will evaluate to a real number between 0 and 1 arbitrary units (au), inclusive, with 0 au indicating that the spectra share no common peaks, and the value 1 au indicating that the spectra are identical. A refinement of cosine similarity is the *identity match factor,*[17] based on the *composite score* described in the seminal paper by Stein and Scott.[11] The identity match factor considers the ratio of relative intensities of adjacent peaks when estimating similarity, thus capturing subtle information about isotopic patterns that is not necessarily reflected in cosine similarity. In most software implementations, identity match factors have been scaled to 100 (e.g., AMDIS[18]) or 999 (e.g., NIST MS Search[19]) and are reported as integers; all similarity scores discussed in this manuscript remain unscaled values between 0 and 1 au.

While not explicitly recommended[b], we can use a threshold similarity score for deciding whether two mass spectra are measurements of different compounds — a task we refer to as negative confirmation in this manuscript. For example, if the identity match factor between the mass spectrum of a case sample and the spectrum of a reference standard is 0.3 au, the case sample is unlikely to be the same compound as the reference. Formally, we can think of this process in terms of binary classification and define the *similarity score test* as

$$\text{similarity score test }(M, \tau_M) = \begin{cases} 0, & M < \tau_M, \\ 1, & M \geq \tau_M, \end{cases}$$

where M is a similarity score between two spectra, τ_M is a threshold similarity score, and class prediction 0 implies that the spectra are measurements of different compounds. Class prediction 1 implies that the spectra are measurements of the same compound, but we know that confirming

[b] In general, mass spectra should be visually inspected to confirm high (or low) similarity scores are not due to measurement inconsistencies (e.g., mass range) or numerical artifacts (e.g., many matching noise peaks contributing to high similarity scores). See Figure 1 for a specific example.

whether two samples are the same compound (positive confirmation) with mass spectral comparisons alone is problematic as discussed later in the manuscript. The challenge with implementing the similarity score test for negative confirmation is that there is no obvious choice for a threshold value, especially not one that is universal across all classes of drugs and all varieties of similarity scores; equivalent similarity scores could be computed in very different situations (Figure 1). Algorithms and numerical approaches that produce counter-intuitive results and require analysts to make critical decisions are unappealing in forensics applications.[20]

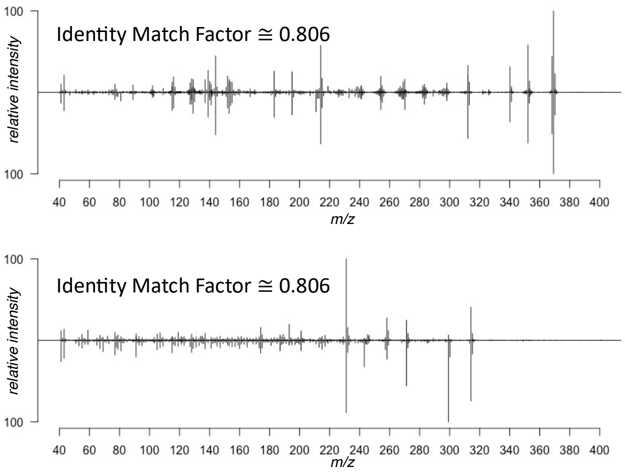

Figure 1: Example of two mass spectral comparisons demonstrating the difficulty interpreting similarity scores without visual inspection. The top example is a head-to-tail plot of two replicate electron ionization mass spectra of the same compound. The bottom example is a head-to-tail plot of two spectra of different compounds. Both pairs of mass spectra had equivalent identity match factors to the third decimal place.

The min-max test was initially formulated during our recent work developing targeted gas chromatography-mass spectrometry methods for identifying synthetic cannabinoids[21] and cathinones.[22] We needed to determine if we could discriminate several pairs of closely eluting compounds based on their mass spectra. We wanted to avoid the ambiguity of threshold setting with a similarity score test and the subjectivity of visual comparison with manual interpretation. The min-max test was able to meet these requirements by leveraging replicate mass spectra to characterize spectral self-similarity within a sample and effectively remove subjectivity from the analysis. In this manuscript, we present the method more completely, with updated notation and refinements to reflect what we learned from our initial experimentation. We discuss how the min-max test, by construction, has an intuitive interpretation for deciding whether spectra are measures of different compounds (negative confirmation), and demonstrate how it out-performs the similarity score test for general classification using EI mass spectra of assorted drugs.

2. Methods

2.1 The min-max test

At the core of the min-max test are replicate measurements. By computing similarity scores between replicate mass spectra of individual compounds, we have context for decisions about pairs of compounds. Assume we have two samples to compare: the first is an unidentified compound isolated from seized evidence, and the second is a standard reference compound. Let $\boldsymbol{S_{11}}$ and $\boldsymbol{S_{22}}$ be sets of intra-sample similarity scores computed between two or more replicate mass spectra of samples 1 and 2, respectively. Let $\boldsymbol{S_{12}}$ be the set of inter-sample similarity scores computed between the spectra of the two samples. Using these sets of values, we can formulate a *spectral comparison index* (δ_i) that follows the general form

$$\delta_i = f\big(g(\boldsymbol{S_{11}}), g(\boldsymbol{S_{22}})\big) - h(\boldsymbol{S_{12}}), \tag{1}$$

where $f(\cdot)$, $g(\cdot)$, and $h(\cdot)$ are functions chosen to reduce the input sets of data into single representative values. In the min-max test, we decide whether two samples are different compounds by comparing the most conservative estimate of intra-sample spectral similarity to the most generous estimate of inter-sample similarity. Formally, we compute the *min-max index* (δ_{MM}) as

$$\delta_{MM} = \min(\min(\boldsymbol{S_{11}}), \min(\boldsymbol{S_{22}})) - \max(\boldsymbol{S_{12}}), \tag{2}$$

where the functions $\min(\cdot)$ and $\max(\cdot)$ denotes the minimum and maximum values contained in the specified sets, respectively. For ease of reading, we will drop the subscript from δ_{MM} and refer to the min-max index with the symbol δ as no other spectral comparison indices are computed in this manuscript.

As presented in Eq. (2) and with score sets $\boldsymbol{S_{11}}$, $\boldsymbol{S_{22}}$, and $\boldsymbol{S_{12}}$ constructed using similarity scores that evaluate between 0 and 1 au, the evaluated δ will be a real number between -1 and 1 au with practical values ranging between -0.1 and 0.9 au. The most intuitive employment of the min-max index is to assess whether $\delta > 0$ au and infer that the compared sets of spectra are measures of different compounds; the larger the value of δ, the more certain we are of the claim. If $\delta \leq 0$ au, there is at least some overlap in the observed intra and inter-sample spectral similarity and so we cannot confidently claim the samples are different compounds.

A transformation such as $\delta' = 1 - \max(0, \delta)$ allows us to compare min-max indices more readily to similarity scores. The transformed min-max index evaluates between 0 and 1 au with uncertainty of a negative confirmation increasing with index values due to increased spectral similarity, as is the case with similarity scores. We can define the *min-max test* as

$$\text{min-max test }(\delta', \tau_{\delta'}) = \begin{cases} 0, & \delta' < \tau_{\delta'}, \\ 1, & \delta' \geq \tau_{\delta'}, \end{cases}$$

where $\tau_{\delta'}$ is a min-max index threshold value, and prediction classes 0 and 1 implying that the spectra are measures of different compounds and the same compound, respectively. An intuitive employment of the min-max test in this formulation is to set $\tau_{\delta'} = 1$ au. For the remainder of this manuscript, all discussion of min-max indices will reference these transformed values.

2.2 Evaluation methodology

To demonstrate and evaluate the min-max test, we collated previously published and newly measured mass spectra into a single collection. The collection contained 10 replicate measurements of 144 illicit drug standards (comprised of synthetic cannabinoids, cathinones and opioids), totaling 1440 mass spectra labeled with names and molecular formulae. With these mass spectra, we computed two *min-max datasets*, the first with min-max indices computed using cosine similarity as the representative similarity score, and the second using identity match factors. These datasets were constructed by computing the min-max indices between all possible pairs of compounds using 3 randomly selected replicate spectra, repeating the experiment 100 times. In cases where the same compounds were being compared, we ensured the selected replicates were not overlapping. The resulting datasets contained 2073600 min-max indices each, approximately 0.7 % of which were computed between the same compound and the rest computed between different compounds. We also generated two *similarity score datasets* to evaluate the similarity score test for comparison. Each of these datasets were constructed by computing all possible non-trivial cosine similarity scores and identity match factors between the 1440 spectra, thus each dataset contained 2072160 total scores. Approximately 0.6 % of the similarity scores in the datasets were computed between spectra of the same compound and the rest between different compounds. The raw mass spectra, computed datasets, and source code and scripts used to analyze results are available for review.[23]

Performance measures: For convenience, we use p to denote an index and τ_p to denote a threshold value associated with p. A positive prediction is when $p \geq \tau_p$, and a negative prediction is when $p < \tau_p$. For a set of indices that can be mapped to binary classifications, the number of True Positives (TP) is the count of positive predictions associated with positive classification (i.e., the compared spectra were replicates of the same compound). True Negatives (TN) are the negative predictions associated with negative classification (i.e., the compared spectra were measurements of different compound). False Positives (FP) are positive predictions that should have been associated with a negative classification, and False Negative (FN) are the negative predictions that should have been associated with positive classifications. Several standard performance measures can be derived from these quantities. In this manuscript, we considered accuracy, true positive rate

(TPR) or recall, specificity, precision, and false positive rate (FPR) as described by Fawcett[24] and summarized in Table 1.

Table 1: Summary of performance metrics considered in this evaluation. TP is the number of correct positive predictions, TN is the number of correct negative predictions, FP is the number of incorrect positive predictions, and FN is the number of incorrect negative predictions.

Metric	**Definition**
Accuracy	$\frac{TP + TN}{TP + FN + FP + TN}$
True Positive Rate (TPR)	$\frac{TP}{TP + FN}$
Specificity	$\frac{TN}{FP + TN}$
Precision	$\frac{TP}{TP + FP}$
False Positive Rate (FPR)	$\frac{FP}{TN + FP}$

Determining optimal thresholds: There are several options for determining optimal threshold values. One approach is to simply select the threshold value that optimizes the objective function (e.g., maximizes test accuracy) for the entire available dataset. While this method is likely to identify a unique threshold for each objective function, we gain no insights about how the identified threshold will perform with new data. A second approach is to use several subsets of the data to determine a range of threshold values, removing some data dependency and shedding some light on what an ideal threshold might look like for completely new data. In this manuscript, we obtained optimal thresholds using complete datasets and an iterative subset selection approach. For iteratives subsets, we randomly selected 10000 indices, ensuring that exactly 60 were values from replicate mass spectra, and repeated the process 1000 times. We determined threshold indices that maximized (1) accuracy, and (2) the difference between recall (TPR) and FPR.

3. Numerical Results

To give a general overview of how similarity scores and min-max indices are distributed across our datasets, we generated box and whisker plots with indices distinguished as estimates of either different compounds or the same compound (Figure 2). Each box and whisker object describes the distribution of the specified index. Outlined boxes display the computed Inter Quartile Range (IQR), with the 2nd quartile (median) marked as a darkened line within the box, and the bottom and top edges of the box indicating the 1st and 3rd quartile indices, respectively. The upper whisker attached to each box indicates the maximum index within 1.5 IQR of the third quartile value, and the lower whisker is the minimum index within 1.5 IQR of the first quartile value. Indices greater than 1.5 IQR of the 1st or 3rd quartile are displayed as outliers (open circles) in the plot. We note that these data points are not outliers in the classical sense, as they do represent acceptable indices computed between pairs and sets of spectra.

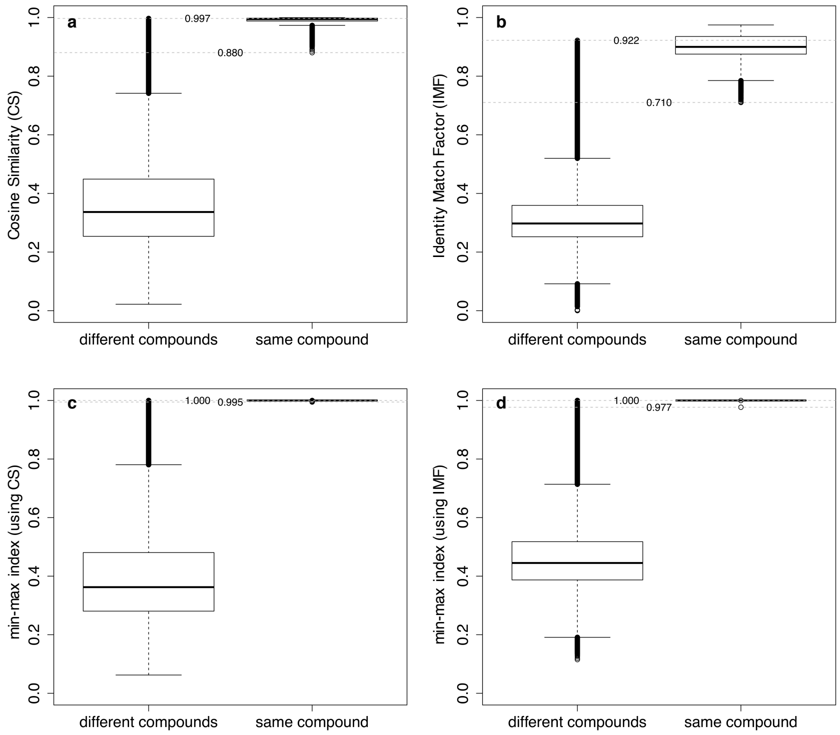

Figure 2: Box and whisker plots summarizing the distribution of (a) cosine similarity scores, (b) identity match factors, (c) min-max indices computed using cosine similarities, and (d) min-max indices computed using identity match factors. Dashed horizontal lines indicate the range of overlapping indices computed between spectra of different and same compounds.

The boxplots clearly illustrate the impact of similarity score selection, cosine similarity (Fig 2a) or identity match factor (Fig 2b), on the constructed similarity score datasets, even though both use the exact same underlying mass spectra. For example, with a threshold of 0.9 au, the accuracy of the similarity score test using the cosine similarities dataset (Fig 2a) is over 98 % with 99 %

recall. The observed accuracy of the similarity score test using the identity match factors dataset (Fig 2b) and a threshold of 0.9 au is approaching 100 %, but with a recall of only 50 %. With the two min-max datasets (Figs 2c and d), the effect of similarity score selection is more subtle, and the indices computed between replicate measurements of the same compound hover close to the 1 au decision-making threshold we intuitively expected. The accuracy of the min-max test with a threshold of 1 au is approaching 100 % with recall values also just under 100 %, using either dataset (Figs 2c and d). A comprehensive performance assessment of both tests using both similarity scoring selections and several fixed threshold values is provided in the Supplementary Information.

Figures 2a and b also confirm that confident negative confirmations can be made using just the similarity score test. For example, there was not a single false negative (incorrect negative confirmation) using a threshold of 0.7 au with either of the similarity score datasets in this manuscript. However, using such a low threshold is a very inefficient filter. We define the *gray area* as the range of similarity scores that will falsely characterize true negatives as false positives using the similarity score test. The lowest observed cosine similarity score for a true positive in the cosine similarity dataset is 0.88 au. We refer to this value as the lower gray area threshold. There were 370 pairs of different compounds where at least one of the computed cosine similarity scores was greater that then the lower gray zone threshold, and so would be falsely classified as the same compound if the lower gray zone threshold was used as the threshold in a similarity score test. Of these pairs, 100 % would be correctly classified as true negatives using the min-max test with a 1 au threshold. Similar results were observed with the gray area false positives from the similarity score test being correctly classified as true negatives using the min-max test when identity match factors were used to estimate spectral similarity (see Table S5 in Supplemental Information).

The optimal decision-making thresholds for both maximizing the accuracy and maximizing the difference between recall and false positive rates using all combinations of tests and similarity scores are summarized in Table 2. The optimal threshold value for using the min-max test is always around 1 au, regardless of similarity estimate and objective function. There is at least one instance where the threshold that maximized the difference between recall and false positive rate was 0.975 au. Because we used randomly selected subsets of data, we cannot easily disaggregate the exact conditions that led to that threshold value. The optimal threshold values for the similarity score tests varied more substantially with choice of similarity estimate and objective, but the tests performed well. That said, the objective values at the optimal threshold with the similarity score test were always less than the objective values using the min-max test with intuitive threshold 1 au.

Table 2: Summary of optimal threshold values identified for maximizing test accuracy and difference between true positive rate (TPR) and false positive rate (FPR) using all available data and randomly selected subsets of data. The first value in each cell lists the optimal threshold for all available data followed by the objective value in parenthesis. The second line in each cell lists the range of values obtained from the randomized subsets (n = 1,000).

Goal:	**Maximizing Accuracy**		**Maximizing TPR-FPR**	
	Similarity Score Test	Min-Max Test	Similarity Score Test	Min-Max Test
Cosine Similarity	0.984 (0.998) 0.971 to 0.994	1 (0.999) 0.996 to 1	0.902 (0.987) 0.880 to 0.982	0.995 (0.998) 0.993 to 1
Identity Match Factor	0.881 (0.997) 0.847 to 0.926	1 (0.999) 1	0.742 (0.984) 0.710 to 0.851	0.999 (0.998) 0.975 to 1

4. Discussion

To begin, we must acknowledge that our seemingly large datasets, with over 2 million data points each, represent a small fraction of the potential chemical space one might explore using EI mass spectrometry – we only considered mass spectra of 144 synthetic cannabinoids, cathinones and opioids. Additionally, our manuscript only evaluated the effect of two different similarity score choices. That said, the numerical results presented in the previous section and as supplemental information support our notion that the min-max test can be an excellent method for objective negative confirmation.

The original motivation for developing the min-max test was to overcome the known limitation of the similarity score test – a good decision-making threshold is difficult to select. While using cosine similarities and identity match factors to estimate spectral similarity is mature in its application, the rough reproducibility of mass spectra and non-linearity of the computations underlying these similarity estimates make them difficult to interpret without visual appraisal, hence limiting their objectivity. In our study, we were able to identify a range of optimal thresholds to maximize similarity score test accuracy and the difference between recall and false positive rate using both cosine similarities and identity match factors. These performance results give us a new data-driven basis from which to select decision-making thresholds for similarity score tests going-forward, yet these values lack the type of intuitive grounding necessary to be completely satisfying. And as with any quantity derived from data, there is concern that the dataset used to determine optimal thresholds was inadequate or inappropriate for the next application of the similarity score test.

Selecting a threshold value for the min-max test is intuitive. A min-max index of 1 au means that the minimum self-similarity observed within the sets of replicate spectra is equal to the maximum similarity observed between the spectra from either set, implying that the sets of spectra are not discernible. The performance results for the min-max test using the full dataset and the threshold of 1 au were excellent, with accuracy, recall and specificity all at least 99.8 %, regardless of similarity score imposed on the method. Using a slightly more conservative threshold value of 0.977 au, we have slightly worse accuracy and specificity, but observe 100 % recall.

The obvious limitation of the min-max test is that it requires replicate mass spectra of each of the compared samples. In an application like seized drug analysis, this requirement is likely not an issue as there is often enough seized evidence and reference standard to take several replicate measurements. However, it is not impossible that a lack of material or analysis time make taking replicate measurements impractical. One useful strategy in these scenarios is to run the similarity score test first, followed by the min-max test if the result falls within the gray area as demonstrated in the numerical results. The other more subtle limitation of the min-max test is that poor quality spectra (e.g., spectra containing contaminants, or measured to inconsistent mass limits) will produce unreliable results. The effectiveness of the min-max test stems from our ability to quantify the conceptual notion of self-similarity. If a poor replicate is included in the test, our understanding of self-similarity is misrepresented in the min-max index, and the utility of the min-max test is essentially void. Alternative spectral comparison indices can be formulated using the general framework in Eq. (1) that are *less* susceptible to failing when provided a single incorrect measurement (e.g., choosing function $g(\cdot)$ to select the median value contained in the set). However, these alternate function choices may lack the simple and intuitive justifications that make the min-max test so satisfying. The similarity score test is also susceptible to failure if provided with poor quality mass spectra; this limitation is far more self-evident and is presently defenseless.

Our manuscript focused on the utility of the min-max test for negative confirmation tasks. Positive confirmation with mass spectrometry is appreciably more difficult. The performance metric of interest for positive confirmation tasks would be precision; the fraction of positive predictions that are true positives. With our datasets and a threshold of 1 au, the precision of the min-max test precision was greater than 85 % (using either cosine similarity or identity match factors) with nearly 100 % recall. This means that up to 15 % of positive predictions were false positives. Given that our dataset consists of several pairs of isomeric drugs, with near identical mass spectra, this high false positive rate is not surprising. A well-known strategy for enhancing the value of similarity scores is to combine them with measures from orthogonal technologies such as retention indices obtained from gas chromatography.[25–27] Exploring how we can effectively combine retention indices within on our framework of spectral comparison indices using replicate measurements for the purposes of objective positive confirmation is on-going work in our lab.

5. Conclusions

Forensic chemists are routinely required to make consequential decisions as quickly as possible. One example of a major decision is confirming that seized evidence does not contain an illegal drug (negative confirmation). The traditional method employed for this task is gas chromatography mass spectrometry followed by mass spectral interpretation. While effective, this approach is inherently subjective. An alternative is to compute numerical estimates of mass spectral similarity and use these values to make decisions. Unfortunately, setting decision making thresholds is a non-trivial challenge, especially across various classes of drugs. In this manuscript, we introduced the min-max test as an alternative method for negative confirmation using mass spectra. We discussed how selecting a threshold for the min-max test is intuitive and demonstrated that the method outperforms automated decision-making using spectral similarity estimates and threshold values, even when the threshold has been optimally chosen for a fixed dataset.

In addition to being effective, algorithms and software tools must be objective and explainable to be of any practical use in a forensic science setting. We constructed the min-max test with these two considerations in mind. We believe the min-max test will be an indispensable tool for forensic chemists performing negative confirmation tasks using mass spectra, and that this manuscript provides a template for further developing objective and explainable methods in the forensic sciences.

6. Acknowledgements

The authors would like to thank Drs. Kearsley, Mallard, Stein and Wallace of the National Institute of Standards and Technology for sharing insights that were instrumental in developing the min-max test. The authors would also like to thank Amber Burns of the Maryland State Police Forensic Sciences Division for sharing practical perspectives on analytical challenges faced by forensic laboratories.

7. Author Contributions

A.S.M. and E.S. conceived the methods and designed the research plan. E.S. conducted laboratory experiments. A.S.M. performed computational analyses. A.S.M. and E.S. co-wrote the manuscript.

8. Competing Interests

The authors declare no competing interests.

9. 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.

10. References

(1) Chauhan, A. GC-MS Technique and Its Analytical Applications in Science and Technology. *J Anal Bioanal Tech* **2014**, *5* (6). https://doi.org/10.4172/2155-9872.1000222.

(2) Biemann, K. *Mass Spectrometry: Organic Chemical Applications*; McGraw-Hill, 1962.

(3) McLafferty, F. W.; Tureček, F. *Interpretation of Mass Spectra*; University science books, 1993.

(4) Gross, J. H. *Mass Spectrometry: A Textbook*; Springer Science & Business Media, 2006.

(5) NIST Mass Spectrometry Data Center. NIST20 Mass Spectral Library (accessed 2021 -04 - 19).

(6) SWGDRUG MS Library Version 3.9 https://swgdrug.org (accessed 2021 -04 -19).

(7) Mayorov, A.; Mirokhin, Y.; Tchekhovskoi, D.; Stein, S. New Developments in the Modelling of Ion Fragmentation by MS Interpret Software. In *Proceedings of the 67th Annual ASMS Conference on Mass Spectrometry and Allied Topics*; Atlanta, Geogia, USA, 2019.

(8) Moorthy, A. S.; Kearsley, A. J.; Mallard, W. G.; Wallace, W. E. Mass Spectral Similarity Mapping Applied to Fentanyl Analogs. *Forensic Chemistry* **2020**, *19*, 100237. https://doi.org/10.1016/j.forc.2020.100237.

(9) Sparkman, O. D. Evaluating Electron Ionization Mass Spectral Library Search Results. *J. Am. Soc. Mass Spectrom.* **1996**, *7* (4), 313–318. https://doi.org/10.1021/jasms.8b00869.

(10) Pesyna, G. M.; Venkataraghavan, Rengachari.; Dayringer, H. E.; McLafferty, F. W. Probability Based Matching System Using a Large Collection of Reference Mass Spectra. *Anal. Chem.* **1976**, *48* (9), 1362–1368. https://doi.org/10.1021/ac50003a026.

(11) Stein, S. E.; Scott, D. R. Optimization and Testing of Mass Spectral Library Search Algorithms for Compound Identification. *J Am Soc Mass Spectrom* **1994**, *5* (9), 859–866. https://doi.org/10.1016/1044-0305(94)87009-8.

(12) Wan, K. X.; Vidavsky, I.; Gross, M. L. Comparing Similar Spectra: From Similarity Index to Spectral Contrast Angle. *J Am Soc Mass Spectrom* **2002**, *13* (1), 85–88. https://doi.org/10.1016/S1044-0305(01)00327-0.

(13) Koo, I.; Zhang, X.; Kim, S. Wavelet- and Fourier-Transform-Based Spectrum Similarity Approaches to Compound Identification in Gas Chromatography/Mass Spectrometry. *Anal. Chem.* **2011**, *83* (14), 5631–5638. https://doi.org/10.1021/ac200740w.

(14) Kim, S.; Koo, I.; Jeong, J.; Wu, S.; Shi, X.; Zhang, X. Compound Identification Using Partial and Semipartial Correlations for Gas Chromatography–Mass Spectrometry Data. *Anal. Chem.* **2012**, *84* (15), 6477–6487. https://doi.org/10.1021/ac301350n.

(15) Garg, N.; Kapono, C. A.; Lim, Y. W.; Koyama, N.; Vermeij, M. J. A.; Conrad, D.; Rohwer, F.; Dorrestein, P. C. Mass Spectral Similarity for Untargeted Metabolomics Data Analysis of Complex Mixtures. *International Journal of Mass Spectrometry* **2015**, 9.

(16) Moorthy, A. S.; Wallace, W. E.; Kearsley, A. J.; Tchekhovskoi, D. V.; Stein, S. E. Combining Fragment-Ion and Neutral-Loss Matching during Mass Spectral Library Searching: A New General Purpose Algorithm Applicable to Illicit Drug Identification https://pubs.acs.org/doi/pdf/10.1021/acs.analchem.7b03320 (accessed 2020 -05 -06). https://doi.org/10.1021/acs.analchem.7b03320.

(17) Moorthy, A. S.; Kearsley, A. J. Pattern Similarity Measures Applied to Mass Spectra. In *Progress in Industrial Mathematics: Success stories*; ICIAM 2019 SEMA SIMAI Springer Series; Springer International Publishing, 2021; Vol. 5, pp 43–54.

(18) D’Arcy, P.; Mallard, W. G. AMDIS–User Guide. *US Department of Commerce, Technology Administration, National Institute of Standards and Technology (NIST), Gaithersburg, MD, USA.[Google Scholar]* **2004**.

(19) Stein, S. E. NIST MS Search v.2.3 chemdata.nist.gov (accessed 2021 -02 -06).

(20) Swofford, H.; Champod, C. Implementation of Algorithms in Pattern & Impression Evidence: A Responsible and Practical Roadmap. *Forensic Science International: Synergy* **2021**, *3*, 100142. https://doi.org/10.1016/j.fsisyn.2021.100142.

(21) Sisco, E.; Burns, A.; Moorthy, A. S. A Framework for the Development of Targeted Gas Chromatography Mass Spectrometry (GC-MS) Methods: Synthetic Cannabinoids. https://doi.org/10.1111/1556-4029.14775.

(22) Sisco, E.; Burns, A.; Moorthy, A. S. Development and Evaluation of a Synthetic Cathinone Targeted Gas Chromatography Mass Spectrometry (GC-MS) Method. https://doi.org/10.1111/1556-4029.14789.

(23) Moorthy, A. S.; Sisco, E. Supplemental data and source code for min-max test research https://data.nist.gov/od/id/mds2-2418 (accessed 2021 -06 -01).

(24) Fawcett, T. An Introduction to ROC Analysis. *Pattern Recognition Letters* **2006**, *27* (8), 861–874. https://doi.org/10.1016/j.patrec.2005.10.010.

(25) Wei, X.; Koo, I.; Kim, S.; Zhang, X. Compound Identification in GC-MS by Simultaneously Evaluating the Mass Spectrum and Retention Index. *Analyst* **2014**, *139* (10), 2507–2514. https://doi.org/10.1039/C3AN02171H.

(26) Halket, J. M.; Przyborowska, A.; Stein, S. E.; Mallard, W. G.; Down, S.; Chalmers, R. A. Deconvolution Gas Chromatography/Mass Spectrometry of Urinary Organic Acids – Potential for Pattern Recognition and Automated Identification of Metabolic Disorders. *Rapid Communications in Mass Spectrometry* **1999**, *13* (4), 279–284. https://doi.org/10.1002/(SICI)1097-0231(19990228)13:4<279::AID-RCM478>3.0.CO;2-I.

(27) Stein, S. E. An Integrated Method for Spectrum Extraction and Compound Identification from Gas Chromatography/Mass Spectrometry Data. *J Am Soc Mass Spectrom* **1999**, *10* (8), 770–781. https://doi.org/10.1016/S1044-0305(99)00047-1.

Supplementary Information for

The min-max test: an objective method for discriminating mass spectra

Arun S. Moorthy and Edward Sisco

National Institute of Standards and Technology, Gaithersburg, MD, USA

The tables S1 through S4 summarize the performance of both the match factor and min-max tests at fixed threshold values, using either (1) all available data (resulting in single performance measures per threshold) or (2) subsets of 10,000 randomly selected indices, repeated 1,000 times. Random selection was constrained such that exactly 60 of the indices represented measurements of the same compound. Threshold values were selected based on visual observation of the boxplots illustrated in Figure 1 of the main manuscript.

Table S5 summarizes findings for how pairs of compounds with spectral similarity scores in the "gray area" for the similarity score test faired using the min-max test.

Table S1: Summary of similarity score test performance at prescribed threshold values when using all available data. Mass spectral similarity was estimated either using cosine similarities or identity match factors. Values with 1 significant digit indicate no rounding. *TPR = True Positive Rate. **FPR = False Positive Rate.

Performance of similarity score test using complete match factor dataset (cosine similarity)

Threshold	**Accuracy**	**TPR***	**Specificity**	**Precision**	**FPR****
0.880	0.984	1	0.984	0.283	0.016
0.900	0.987	0.999	0.987	0.329	0.013
0.920	0.991	0.994	0.991	0.400	0.009
0.940	0.994	0.983	0.994	0.515	0.006
0.960	0.995	0.947	0.996	0.611	0.004
0.980	0.997	0.890	0.999	0.792	0.001
0.988	0.995	0.245	1	1	0

Performance of similarity score test using full match factor dataset (identity match factors)

Threshold	**Accuracy**	**TPR***	**Specificity**	**Precision**	**FPR****
0.650	0.971	1	0.970	0.177	0.029
0.700	0.981	1	0.981	0.250	0.019
0.750	0.988	0.994	0.988	0.349	0.012
0.800	0.994	0.949	0.994	0.495	0.006
0.850	0.997	0.877	0.997	0.682	0.003
0.900	0.997	0.499	1.000	0.933	0.000
0.950	0.994	0.078	1	1	0

Table S3: Summary of min-max test performance at prescribed threshold values when using all available data. Mass spectral similarity was estimated either using cosine similarities or identity match factors. Values with 1 significant digit indicate no rounding. *TPR = True Positive Rate. **FPR = False Positive Rate.

Performance of min-max test using full min-max dataset (cosine similarity)

Threshold	**Accuracy**	**TPR***	**Specificity**	**Precision**	**FPR****
0.970	0.995	1	0.995	0.590	0.005
0.975	0.996	1	0.996	0.636	0.004
0.980	0.997	1	0.997	0.687	0.003
0.985	0.997	1	0.997	0.732	0.003
0.990	0.998	1	0.998	0.775	0.002
0.995	0.999	1.000	0.999	0.865	0.001
1.000	0.999	0.998	0.999	0.908	0.001

Performance of min-max test using full min-max dataset (identity match factors)

Threshold	**Accuracy**	**TPR***	**Specificity**	**Precision**	**FPR****
0.970	0.996	1	0.996	0.663	0.004
0.975	0.997	1	0.997	0.693	0.003
0.980	0.997	1.000	0.997	0.730	0.003
0.985	0.998	1.000	0.998	0.763	0.002
0.990	0.998	1.000	0.998	0.792	0.002
0.995	0.999	1.000	0.998	0.822	0.002
1.000	0.999	1.000	0.999	0.852	0.001

*Table S2: Summary of similarity score test performance at prescribed threshold values when using randomly selected sets of 10,000 indices, repeated 1000 times. Sets were selected such that exactly 60 indices were estimates between replicate mass spectra of the same compound. Mass spectral similarity was estimated using either cosine similarities or identity match factors. Values with 1 significant digit indicate no rounding. *TPR = True Positive Rate. **FPR = False Positive Rate.*

Performance of the similarity score test using random subsets of the match-factor dataset (cosine similarity)

Threshold	Accuracy	TPR*	Specificity	Precision	FPR**
0.880	0.980 to 0.988	1	0.980 to 0.988	0.229 to 0.330	0.012 to 0.020
0.900	0.983 to 0.991	0.967 to 1	0.983 to 0.991	0.264 to 0.403	0.009 to 0.017
0.920	0.988 to 0.993	0.950 to 1	0.988 to 0.993	0.326 to 0.472	0.007 to 0.012
0.940	0.991 to 0.996	0.917 to 1	0.991 to 0.996	0.410 to 0.615	0.004 to 0.009
0.960	0.994 to 0.998	0.833 to 1	0.994 to 0.998	0.487 to 0.740	0.002 to 0.006
0.980	0.996 to 0.999	0.733 to 0.983	0.997 to 0.999	0.638 to 0.914	0.001 to 0.003
0.998	0.995 to 0.997	0.083 to 0.483	1	1	0

Performance of the similarity score test using random subsets of the match-factor dataset (identity match factor)

Threshold	Accuracy	TPR*	Specificity	Precision	FPR**
0.650	0.965 to 0.977	1	0.965 to 0.976	0.147 to 0.203	0.024 to 0.035
0.700	0.977 to 0.985	1	0.977 to 0.985	0.207 to 0.290	0.015 to 0.022
0.750	0.985 to 0.992	0.950 to 1	0.985 to 0.992	0.282 to 0.435	0.008 to 0.015
0.800	0.991 to 0.996	0.850 to 1	0.992 to 0.996	0.403 to 0.584	0.004 to 0.008
0.850	0.995 to 0.998	0.750 to 0.993	0.996 to 0.999	0.558 to 0.821	0.001 to 0.004
0.900	0.995 to 0.998	0.267 to 0.733	0.999 to 1	0.667 to 1	0 to 0.001
0.950	0.994 to 0.995	0 to 0.200	1	1	0

*Table S4: Summary of min-max test performance at prescribed threshold values when using randomly selected sets of 10,000 indices, repeated 1000 times. Sets were selected such that exactly 60 indices were estimates between replicate mass spectra of the same compound. Mass spectral similarity was estimated using either cosine similarities or identity match factors. Values with 1 significant digit indicate no rounding. *TPR = True Positive Rate. **FPR = False Positive Rate.*

Performance of min-max test using random subsets of the min-max dataset (cosine similarity)

Threshold	Accuracy	TPR*	Specificity	Precision	FPR**
0.970	0.993 to 0.998	1	0.993 to 0.998	0.444 to 0.714	0.002 to 0.008
0.975	0.994 to 0.998	1	0.994 to 0.998	0.483 to 0.769	0.002 to 0.006
0.980	0.995 to 0.999	1	0.995 to 0.999	0.530 to 0.822	0.001 to 0.005
0.985	0.996 to 0.999	1	0.996 to 0.999	0.594 to 0.870	0.001 to 0.004
0.990	0.997 to 0.999	1	0.996 to 0.999	0.632 to 0.896	0.001 to 0.004
0.995	0.998 to 1.000	0.983 to 1	0.998 to 1.000	0.732 to 0.968	0.000 to 0.002
1.000	0.998 to 1	0.967 to 1	0.998 to 1	0.779 to 1	0 to 0.002

Performance of min-max test using random subsets of the min-max dataset (identity match factor

Threshold	Accuracy	TPR*	Specificity	Precision	FPR**
0.970	0.995 to 0.999	1	0.995 to 0.998	0.536 to 0.800	0.002 to 0.005
0.975	0.995 to 0.999	1	0.995 to 0.999	0.556 to 0.811	0.001 to 0.005
0.980	0.996 to 0.999	0.983 to 1	0.996 to 0.999	0.600 to 0.857	0.001 to 0.004
0.985	0.996 to 0.999	0.983 to 1	0.996 to 0.999	0.625 to 0.882	0.001 to 0.004
0.990	0.997 to 0.999	0.983 to 1	0.997 to 0.999	0.652 to 0.909	0.001 to 0.003
0.995	0.997 to 1.000	0.983 to 1	0.997 to 1.000	0.690 to 0.938	0.000 to 0.003
1.000	0.997 to 1.000	0.983 to 1	0.998 to 1.000	0.723 to 0.968	0.000 to 0.002

Table S5: Summary of false positives from the similarity score test, identified using a lower gray area threshold, that were correctly classified using the min-max test. The cosine similarity score for lower gray area threshold was 0.880 au, resulting in 370 false positives. The identity match factor for lower gray area threshold was 0.710, resulting in 460 false positives.

	Fraction of false positive pairs from similarity score test corrected to true negatives using the min-max test	
min-max test threshold	**Cosine Similarity Dataset**	**Identity Match Factor Dataset**
1	1	0.998
0.995	0.976	0.987
0.985	0.903	0.970
0.975	0.850	0.950