NIST Technical Note
NIST TN 2331

HDCMS

a package for computing high-dimensional consensus mass spectral similarity scores

Jason J. Eveleth
Arun S. Moorthy
Anthony J. Kearsley

This publication is available free of charge from:
https://doi.org/10.6028/NIST.TN.2331

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

NIST Technical Note
NIST TN 2331

HDCMS

a package for computing high-dimensional consensus mass spectral similarity scores

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

Arun S. Moorthy
Department of Forensic Science
Trent University
1600 West Bank Drive, Peterborough, ON, K9L 0G2, Canada

This publication is available free of charge from:
https://doi.org/10.6028/NIST.TN.2331

April 2026

U.S. Department of Commerce
Howard Lutnick, Secretary

National Institute of Standards and Technology
Craig Burkhardt, Acting Under Secretary of Commerce for Standards and Technology and Acting NIST Director

Publication History
Approved by the NIST Editorial Review Board on 2025-02-13

How to cite this NIST Technical Series Publication:
Eveleth JJ, Moorthy AS, Kearsley AJ (2026) HDCMS: a package for computing high-dimensional consensus mass spectral similarity scores. (National Institute of Standards and Technology, Gaithersburg, MD), NIST Technical Note (TN) NIST TN 2331 https://doi.org/10.6028/NIST.TN.2331

Author ORCID iDs
Arun S. Moorthy: 0000-0002-5988-1389
Anthony J. Kearsley: 0000-0002-9576-0621

Contact Information
anthony.kearsley@nist.gov

Abstract

Identification of unknown mass spectra is a fundamental task in analytical chemistry. The unambiguous identification of mass spectra remains a challenge for compounds similar in structure. A novel mathematical approach incorporating measurement variability in replicate mass spectra was proposed in Roberts et al., [1]. The method shows potential in discriminating similar compounds given replicate mass spectra, and has been implemented in a software package called HDCMS (high dimensional consensus mass spectra). In this work, we describe the algorithmic considerations underpinning the HDCMS package, including: data preprocessing, constructing HDCMS from replicate measurements, and comparing pairs of HDCMS. We then demonstrate use of the software and summarize the performance on two lab collected data sets.

Keywords

C programming language; Cheminformatics; Mass spectrometry; Python; Similarity scores.

Table of Contents

List of Tables

List of Figures

Acknowledgments

We are grateful to Dr. Matthew Roberts for his contribution to the development of the algorithms implemented by HDCMS, Drs. Deb McGlynn, Melinda Kleczynski and Brad Alpert for careful readings of and useful feedback to an early version of this report and Dr. Jennifer Bonetti for providing data that was crucial to testing our algorithm.

Author Contributions

Jason J. Eveleth: Methodology, Software, Writing- Original draft preparation. **Arun S. Moorthy**: Conceptualization, Methodology, Data curation, Writing. **Anthony J. Kearsley**: Conceptualization, Methodology, Visualization, Data curation, Writing, Investigation.

1. Introduction

Assigning identifications to unknown analytes given only analyte mass spectral data is an important task in a wide range of fields. In an electron ionization mass spectrum, a compound's measured mass-to-charge (*m/z*) ratios and relative peak intensities are represented as vectors from which similarity can be calculated between the unknown and reference spectra. Commonly used methods for compound identification, include the National Institute of Standards and Technology (NIST) *simple similarity search* (SSS) algorithm and the *identity similarity search* (ISS) algorithm both of which are derived from cosine similarity [2–4]. A comparison of these commonly used methods to other measures of similarity (ℓ_1, ℓ_2, Wasserstein metric) find that cosine similarity derived methods perform better in computing the similarity between replicates of one compound [4]. However, many individual compounds have very similar mass spectral signatures making them difficult to discriminate using either ISS or SSS. Additional information such as the compound retention index [5], compound elution order or visual discrimination by a trained analyst is often required to verify the compound identification [1, 3].

Recently, a new method for comparing and identifying unknown mass spectra called the *high dimensional consensus mass spectral* (HDCMS) algorithm was proposed. The method has been successfully applied to the identification to methamphetamine and phentermine collected using direct analysis real time mass spectrometry (DART-MS) [1] as well as discriminating between structurally similar terpenes [6] among other applications. The HDCMS method considers measurement variability in both mass-to-charge (*m/z*) ratios and relative peak intensities using replicate measurements of mass spectra [1]. The algorithm constructs probability density functions of these two variables from replicate mass spectra for each unknown compound [7]. Incorporating this additional information captures measurement variability and may increase positive identification of an unknown analyte and decrease the need for secondary identification verification. In Roberts, et al., 2022 HDCMS was successfully employed discriminated between methamphetamine and phentermine; two compounds that could not be discriminated by cosine similarity based measures. However, these compounds were not correctly identified by the HDCMS method when instrument parameters varied and sampling order changed [1]. Additionally, the method was employed solely on data collected by DART-MS and can benefit from analysis on data collected using additional mass spectrometry methods. Therefore, this method may improve identification of compounds that are difficult to differentiate using cosine similarity derived comparison methods. However, it requires a flexible implementation of the algorithm to allow for systematic experimentation and to learn of the applications and limitations of the HDCMS approach.

In this work we present a software package called HDCMS that constructs high-dimensional consensus mass spectra and computes the similarity between these data structures. We outline here a complete description of data preprocessing by the HDCMS package, calculation of high dimensional consensus mass spectra, and the computation of a similarity

measure between library and query HDCMS. Additionally, we then demonstrate its performance on two sets of lab collected data sets.

2. Software description

The software, `HDCMS`, is a Python package that wraps a C programming language library. The Python package's source is located here, and the C library's source is located here. The Python package is the main interface for the software. The primary purpose of the software is calculate the similarity of sets of replicate measurements. It relies on two algorithms, xy-hdc and y-hdc explained in the next two sections.

2.1. xy-hdc Algorithm

The mathematical model behind the xy-hdc algorithm was presented in [1]. We reproduce an explanation here for the interested reader; however, knowledge of this is not necessary to use the package. The xy-hdc algorithm takes a list of n replicate measurements and finds the point p with the highest y-value. It finds points that are close to p in each replicate measurement, and computes the mean and standard deviation for mass-to-charge ratio (x) and intensity (y) for those points. Then it marks those points as visited, and repeats this process to until all the points in the replicates have been visited. We will refer to the statistics for each set of points as a Gaussian pdf. The algorithm returns an array of pdf's which is the representation of an xy-hdc summary statistic. A visualization of a xy-hdc summary statistic of the data in Figure 1a gives Figure 1b.

Figure 1b is therefore a top down view of several Gaussian pdfs. As you can see, the set of points on the bottom left form a group with large x and y variance. The string of points in the middle were all grouped together. And finally the tight concentration of points in blue on the right form a third group with small variance.

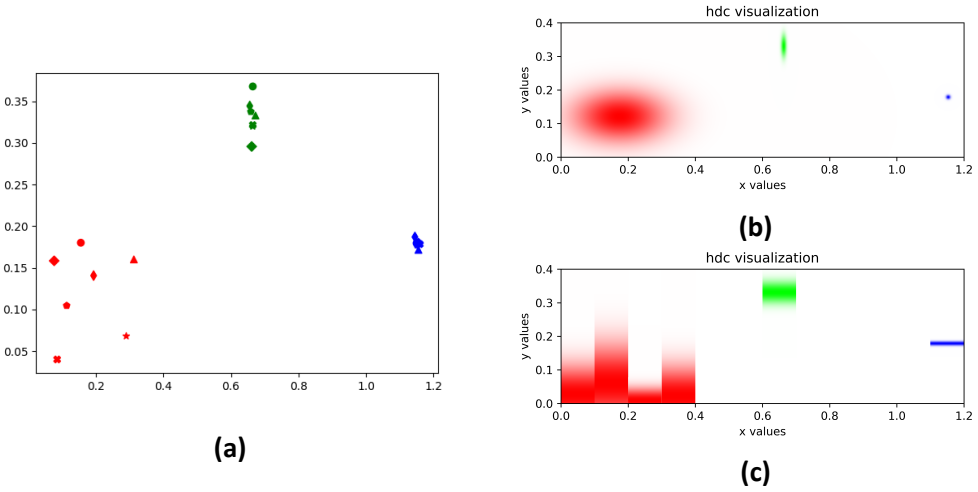
Since our groups of points are Gaussian functions, and since Gaussians are in $L^2(\mathbb{R}^2)$ (square integrable functions of two real variables), we can measure the similarity between two groups of points using the inner product. Thus, we get the similarity by

$$\text{sim}(f,g) = \frac{\langle f,g\rangle}{||f||||g||} = \frac{\langle f,g\rangle}{\sqrt{\langle f,f\rangle}\sqrt{\langle g,g\rangle}}.$$

For details on how this is calculated see [1, 7].

Calculation of the similarity between two xy-hdc summaries is a two step process. We first pick the Gaussian with the largest y-value. We then find a corresponding Gaussian in the other summary, and mark them as visited. Then, we take a weighted average of the similarities of the matched Gaussians. Let S be one xy-hdc summary, and S' be another xy-hdc summary. The resulting similarity is calculated as follows

$$\frac{\sum_{s\in S, s'\in S'} \text{sim}(s,s')\cdot \text{yval}(s)\cdot \text{yval}(s')}{\sum_{s\in S, s'\in S'} \text{yval}(s)\cdot \text{yval}(s')},$$



(a) (b) (c)

Fig. 1. A visualization of the data structures used by HDCMS package on synthetic data. Subfigure 1a shows 7 replicate measurements sampled from 3 imaginary objects (a red object, green object, blue object). The corresponding parts of both summaries are colored according to the imaginary object that they came from. The data was chosen to demonstrate 3 different scenarios, high variance in both directions (red), high y variance and low x variance (green), and low variance in both directions. Subfigure 1b shows the xy-hdc summary for the data. Each object's points are grouped together and represented by a 2D Gaussian pdf. The visualization is top-down. Subfigure 1c shows the y-hdc summary for the data. Subfigure 1c shows several 1D Gaussian pdfs. The points in the bottom left corner were split up into different bins. The points in the middle were grouped into a bin with large variance, and the points on the right were put in a bin with low variance. This is why the assumption that there is small x-variance is important, since binning can split up logically similar items.

with $\text{yval}(s)$ being the mean of the y-values in Gaussian s, and $\text{sim}(s, s')$ being the similarity of the Gaussians as described in the previous paragraph. The similarities are reported from 0 (not similar at all, i.e. the Gaussians in the summaries do not share any area) to 1 (when the Gaussians in the summaries are identical).

2.2. y-hdc Algorithm

The y-hdc algorithm was inspired by the xy-hdc one. The y-hdc algorithm relies on the fact that mass-to-charge ratio usually has much less variability between measurements than intensity. We exploit that by binning the x values for each replicate measurement. This binning does not cause much information loss because the x values are always close to each other for corresponding Gaussians. Then, the summary statistic is the mean and standard deviation (of the y values) for the bins (an array of shape $n \times 2$). These bins are a good proxy for grouping ions based on mass, so this provides a robust fingerprint for the replicate measurements.

We can use an analogous technique to the xy-hdc case and consider each bin to be a one-dimensional Gaussian function. We can compare these Gaussians analogously to the way we did for xy-hdc. For the similarity between a bin represented with mean u_i, variance s_{u_i}, and a bin represented with mean v_i, and variance s_{v_i}, we use the inner product on $L^2(\mathbb{R})$, between two Gaussians with means u_i, v_i, variances s_{u_i}, s_{v_i}. We use that formula in the final similarity like so:

$$\frac{\sum_{i=1}^{n} u_i v_i \text{sim}(u_i, v_i, s_{u_i}, s_{v_i})}{\sum_{i=1}^{n} u_i v_i},$$

which is the weighted average of similarities by the product of the means. A visualization of a y-hdc can be seen in Figure 1.

2.3. The Software Package

The Python package can be installed with pip: `pip install hcdms`. Using the package is a three step process, depicted in Figure 2: (1) preprocess the data to match the format expected by the package library. (2) split replicates into groups and compute a summary statistic for each group. (3) calculate similarity of the summary statistics.

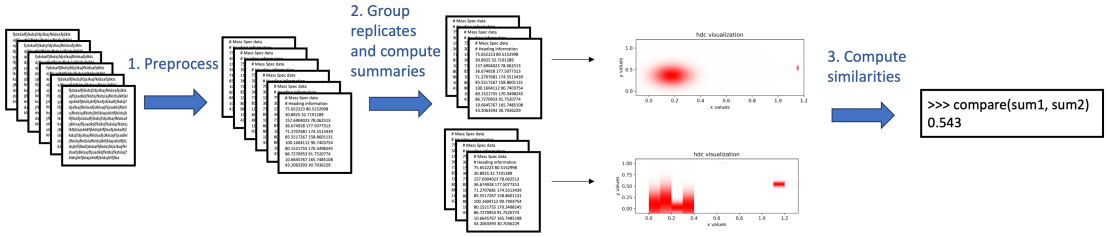


Fig. 2. A depiction of the pipeline used by our software package. The first step is preprocessing the data. More information on data format can be found in Section 2.3. The next step is to group the replicates into summary statistics. This can be done with several data collection functions for more details see the examples.

(1) The data format expected by the C library is straightforward. The format requires distinct replicate measurements to be in separate files. The data must have each mass-to-charge ratio and intensity pair to be on a single line, and separated by whitespace or commas. They can be in normal floating point format or scientific notation. The format permits an optional header with leading # on each line. We provide a verification function `is_ms_valid_data_format` which takes a filename and will raise an exception if the file is incorrectly formatted.

(2) The package provides many data loading routines, which will gather replicate measurements and compute a summary statistic (of either type). Each of them takes in a collection

of filenames (as a list, a regular expression, comma separated string, etc.) and returns a single summary statistic for all the files. If requesting an xy-hdc, an optional `xtol` parameter allows choice of how much x distance is accepted in a Gaussian. The default variance is infinity, which means points arbitrarily far apart can be grouped into a Gaussian. If requesting a y-hdc, an optional `start`, `stop`, `num_bins` can be provided to customize the binning for the summary. The default bins for y-hdc are width 0.1 from values 0 to 900. Binning in this scenario is often harmless as MS data is often provided on the integer m/z value [8]. This can be changed in any call to the y-hdc grouping routines which take in a `start`, `end`, and `num_bins`. Both algorithms allow for a scaling parameter `scale`. This can be 'm' which scales the data so the max value is 1, 'u' for scaling the data to a unit vector, and 'n' which means no scaling. We also provide a visualization routine `write_image` for summaries.

(3) A similarity function, `compare`, is also provided that will automatically detect whether the summary statistics are y-hdc or xy-hdc summaries, and compute the correct comparison. It returns a float if 2 arguments were supplied, or a (symmetric) matrix of floats if more than 2 were supplied. The similarities are reported from 0 (not similar at all) to 1.

The similarity calculation is parameterized by several values. First, we are adding a constant (which defaults to 1e-4) to all of the standard deviations to avoid singularities. This is necessary because a small standard deviation could cause a divide-by-zero-error when calculating the similarity. This parameter can be changed in any call to `compare`. Second, the number of features used in the comparison. By default we use all the features in the summary, but this can be restricted to only n features with largest y-value. This is achieved with the `npeaks` variable passed to `compare`.

3. Illustrative examples

We will show two examples to demonstrate the usage of the Python package. First, we will show a simple use of the xy-hdc algorithm for comparing two sets of replicates, and outline how the pipeline works in practice. Next, we will show how to implement a mass spectral library search with the y-hdc algorithm. Both of these examples are available on github in the examples directory.

3.1. Simple xy-hdc Example

Our first dataset was used by [1]. It is DART-MS data measured at energy level 30V, 60V, and 90V. For this example we only need two compounds, so we will use Cotinine and Serotonin samples. Both compounds have 5 replicate measurements.

The first step, is preprocessing the data. An excerpt from one of the sample files can be seen in Table 1. Note that the header of the file is in the correct form (it is prefixed with #), so it will be ignored. This file is already in the correct format and no further work must be done.

Table 1. The raw centroid data from a DART-MS sample of Cotinine [1]. There is a header, with lines prefixed by #'s, and the data is provided in two columns of floating point numbers. Parts of the file were elided for brevity. This is a valid data format.

```
##JEOL-DX=2.00
##DATA TYPE=Mass Spectrum
##DATA PROCESSING=unrounded
... [elided] ...
##NPOINTS=30
##DATA=
98.061763 1292.655063
146.060204 957.109484
147.093436 723.252767
161.107627 2460.936046
163.123678 3798.993256
... [elided] ...
```

Next, we move on to step 2 of the pipeline: we group the replicates and turn them into summary statistics. The routine for this is `regex2stats2d`, which will take a regular expression matching file names, and returns a xy-hdc summary statistic. The arguments are the regular expression which matches file names, and the directory the files are in. We use an x-tolerance of 0.1 to avoid peaks of high x variance. Next, we can use the visualization function `write_image` to look at one of these data (Figure 3).

Finally, step 3 of the pipeline: we can compare two spectra. We use the function `compare`. This gives a value of 3.58e-07. On its own, this has little meaning. In the next example, we will see how to give context to this value.

3.2. y-hdc Library Search Implementation

Our second dataset is 10 compounds, each with 10 replicate measurements. This data is a subset of the dataset used by [9] and is available publicly here. The data consist of mass spectra of synthetic cannabinoids, cathinones, and opioids.

We implement and use an $ad - hoc$ mass spectral classification library. We load all the replicates for each compound into a list, and use those summaries as a library. Then use that library to identify the other replicates.

We can create the library using the `regex2stats1d` function on each set of replicates (we break up each compound into two sets of 5 replicates). Then, the search function will create a list of names and similarity scores, and sort them. We can use this search function to determine the compounds which are most similar to our query compound. In Table 2, we can see the results of a library search for Acrylfentanyl (C22H26N2O).

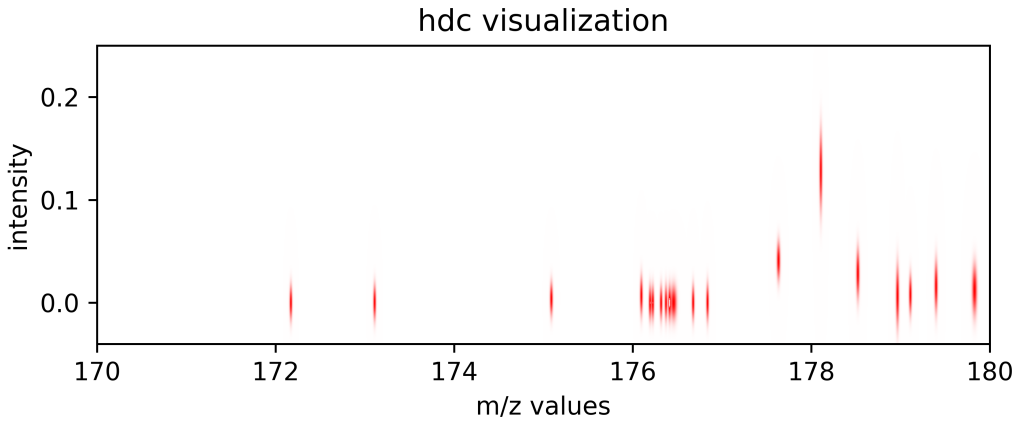


Fig. 3. A visualization of xy-hdc which shows some of the Gaussians of a Cotinine summary statistic. Our image is magnified or zoomed-in in order to better see the Gaussian structure.

Table 2. The results of our library search for Acrylfentanyl (C22H26N2O). The compound names are on the left, and the similarity of the compound and Acrylfentanyl are on the right. Acrylfentanyl had a similarity score of 1.0. This is to be expected since a score of 1.0 means the summary statistics are identical, and one of the compounds we are comparing to is in fact the same summary statistic.

Acrylfentanyl	1.0
4'-methyl Acetyl Fentanyl	0.102
Carfentanil	0.093
Crotonyl Fentanyl	0.09
Cyclopropyl Fentanyl	0.087
p-Fluorobutyryl Fentanyl	0.075
p-Fluorofentanyl	0.073
FIBF	0.057
p-Methoxyfentanyl	0.053
3-Furanyl fentanyl	0.028

4. Impact

Positive identification of unknown compounds is crucial for determining, for example, if a compound is carcinogenic, benign, medicinal, or illicit. With an immense number of isomers or structurally similar compounds that return with high match factors using conventional methods, additional methods are necessary to improve identification of these compounds and decrease the subjectivity in compound identification. HDCMS can be employed on replicate measurements of mass spectra to improve the certainty of compound identification. Here, we employed HDCMS on methamphetamine derivatives and terpene isomers to incorporate measurement uncertainty when determining compound similar-

ity. Unlike other methods, HDCMS requires the collection of replicate measurements for both the library and unknown compound to calculate a probability density function to compare between. However, as shown from the examples, the method outperforms the most commonly employed methods while retaining result interpretability and providing measurement variability. Additionally, due to the methods inherent ability to capture vari-ability, it opens the door for uncertainty quantification in mass spectral measurements and identification.

To date, this method has been employed in the area of chemometrics but its potential applications are vast. The method can be employed in fields with data that is cyclic, and therefore has replicate temporal periods of data(time series)to predict future outcomes. The method could also be used in image analysis to verify similarity between replicate images, for example. Applications beyond these examples certainly exist and this work has only scratched the surface of its potential.

The generality of the software allows `HDCMS` to be used outside the domain of MS. Because of the flexibility of the algorithm, `HDCMS` can be used on an arbitrary x-y data with replicates. This opens up the possibilities of applications to many domains.

With the public release of this software, we hope that the analytical chemistry community and scientific community at large can use `hdmcs` as an easy extension of their existing workflows. The added sensitivity demonstrated here motivates further investigation and contemplation of incorporating replicate measurements in the larger chemometric identi-fication and classification process.

5. Conclusions

In this work we developed a package called HDCMS from a new mathematical approach that calculates similarity between replicate mass spectral measurements. This method incorporates measurement variability of replicate measurements to more holistically cap-ture the similarity between unknown compounds and curated library spectra. Replicate measurements for both unknown and library compounds are required to build probability density functions to calculate similarity. HDCMS was successfully used to positively identify fentanyl analogs and terpene isomers. HDCMS converts replicates of raw text file data to probability density functions. The resulting data structures can then be compared and vi-sualized using the package. A significant advantage of this package is result interpretability when compared to the commonly used NIST mass spectral library. HDCMS returns a ranked hitlist much like NIST's MS Search hitlist. One major advantage of HDCMS is the incorpo-ration of measurement variability which will not only improve mass spectral identification certainty but also allow for the uncertainty quantification in mass spectral measurement and identification.

In addition to a variety of applications testing of this approach including identifying fentanyl and fentanyl analogs [10] and other forensic applications, future work includes the use of

Kolmogorov-Smirnoff machinery [11] to construct probabilistic similarity scores for sets of replicate mass spectra and evaluate intensity differences across peaks. These methods appear to yield accuracy comparable to the HDC scores while requiring fewer replicates though testing is still ongoing. Additionally, comparison with other strategies, [12], may shed light on algorithmic strengths and weaknesses.

6. 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.

7. NIST Disclaimer

 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.

References

[1] Roberts MJ, Moorthy AS, Sisco E, Kearsley AJ (2022) Incorporating measurement variability when comparing sets of high-resolution mass spectra. *Anal Chim Acta* 1230:340247. https://doi.org/https://doi.org/10.1016/j.aca.2022.340247

[2] Stein SE, Scott DR (1994) Sciencedirect - journal of the american society for mass spectrometry : Optimization and testing of mass spectral library search algorithms for compound identification. *Journal of the American Society for Mass Spectrometry* 5:859–866. Available at http://www.sciencedirect.com/science/article/pii/1044030594870098.

[3] Moorthy AS, Wallace WE, Kearsley AJ, Tchekhovskoi DV, Stein SE (2017) Combining fragment-ion and neutral-loss matching during mass spectral library searching: A new general purpose algorithm applicable to illicit drug identification. *Analytical Chemistry* 89:13261–13268. https://doi.org/10.1021/acs.analchem.7b03320

[4] Moorthy AS, Kearsley AJ (2021) Pattern similarity measures applied to mass spectra. *Progress in Industrial Mathematics: Success Stories: The Industry and the Academia Points of View* (Springer), pp 43–53. https://doi.org/https://doi.org/10.1007/978-3-030-61844-5_4

[5] Qu C, Schneider BI, Kearsley AJ, Keyrouz W, Allison TC (2021) Predicting kovats retention indices using graph neural networks. *Journal of Chromatography A* 1646:462100. https://doi.org/https://doi.org/10.1016/j.chroma.2021.462100

[6] McGlynn DF, Rabe Andriamaharavo N, Kearsley AJ (2024) Improved discrimination of mass spectral isomers using the high-dimensional consensus mass spectral similarity algorithm. *Journal of Mass Spectrometry* 59(10):e5084. https://doi.org/https://doi.org/10.1002/jms.5084

[7] Kearsley A, Matthew R (2024) Similarity measures of mass spectra in hilbert spaces. https://doi.org/https://doi.org/10.6028/NIST.TN.2297.

[8] Khrisanfov M, Samokhin A (2022) A general procedure for rounding m/z values in low-resolution mass spectra. *Rapid Commun Mass Spectrom* 36(11):92–94. https://doi.org/https://doi.org/10.1002/rcm.9294

[9] Moorthy AS, Sisco E (2021) The min-max test: An objective method for discriminating mass spectra. *Analytical Chemistry* 93(39):13319–13325. https://doi.org/10.1021/acs.analchem.1c03053

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

[11] Krishnaswamy-Usha A, Capistran BA, Kearsley AJ (2025) Novel probabilistic similarity scores for sets of replicate ei mass spectra. *Analytica Chimica Acta* 1351:343858. https://doi.org/https://doi.org/10.1016/j.aca.2025.343858

[12] Adeoye AI, Jackson GP (2025) Application of the expert algorithm for substance identification (easi) to the electron ionization (EI) mass spectra of fentanyl isomers and analogs. *Forensic Chemistry* :100660https://doi.org/https://doi.org/10.1016/j.forc.2025.100660