Forensic Chemistry 51 (2026) 100795



Contents lists available at ScienceDirect

Forensic Chemistry

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



Clustering and multidimensional scaling of hybrid similarity search results: Application to illicit drug mass spectra

Melinda A. Kleczynski [a,b], Arun S. Moorthy [c], Anthony J. Kearsley [a,*]

[a] National Institute of Standards and Technology, 100 Bureau Drive, Mailstop 8910, Gaithersburg, MD, 20899, USA
[b] Georgetown University, 37th and O Streets, N.W., Washington, DC, 20057, USA
[c] Trent University, 1600 West Bank Drive, Peterborough, ON, K9L 0G2, Canada

ARTICLE INFO

Dataset link: https://github.com/usnistgov/cluster-mass-spec-hybrid-search

Keywords:
Amphetamines
Clustering
Dimensionality reduction
Fentanyl analogs
Mass spectral library searching
Optimization
Peak shifting

ABSTRACT

Multidimensional scaling (MDS) of Hybrid Similarity Search (HSS) Match Factors has been successfully employed to explore mass spectral datasets of compounds with forensic significance, such as potential illicit drugs. Clustering hit lists generated by HSS can reveal groups of compounds such as different types of fentanyl analogs. These clusters can be observed in the original HSS data and in low-dimensional MDS configurations, although this may require alterations to typical analysis pipelines. For some query spectra, changing the loss function used when performing MDS can produce significantly different MDS configurations, particularly with respect to the separation of the point representing the query spectrum from other clusters in the hit list. We show that, in some cases, the SSTRESS loss function is preferable to the commonly used STRESS loss function. In the SSTRESS loss function, the squared residuals are weighted by the squared sum of the dissimilarities between mass spectra and the distances between MDS points. This may be beneficial for clustering applications when the layout of points within a cluster may be less important than the distances between clusters. While SSTRESS may not currently be part of standard mass spectral analysis pipelines, it presents an opportunity for improving MDS utility. Finally, for the hit lists we analyzed, the single linkage clustering results for SSTRESS MDS configurations were highly robust to the choice of random initial configuration, promoting reproducibility in forensic applications.

1. Introduction

The rapid emergence of novel illicit drugs complicates efforts to identify seized compounds [1]. Gas chromatography-mass spectrometry (GC–MS) is frequently used for chemical identification, including for fentanyl and related compounds [2] and amphetamine-type stimulants [3]. Searching reference libraries based on similarity of mass spectra is an established chemical identification technique [4]. Library searching is followed by interrogation of numerical match factors, visual inspection of the mass spectra, and consideration of other contextual information.

Similarity can be quantified through various match factor scores. The match factors considered in this work (simple, hybrid, and symmetric hybrid) have values which range from 0 to 999, with higher values generally indicating better matches. There are rough guidelines presented in the literature for interpreting match factor values for EI mass spectra. Some sources state that values greater than 800 constitute good matches [4,5], while some set this threshold at 900 [6]. It has been stated that matches less than 700 are generally considered questionable [5], while in other circumstances identifications may be appropriate for compounds with matches below 700 [7]. Ultimately, any decision regarding a query spectrum is the responsibility of the analyst [8]. There is work to be done before identification can become a completely objective, numerical task.

Mass spectral similarity can be quantified through a variety of algorithms, having different strengths and resulting in different distributions of similarities [9]. One such algorithm is the Hybrid Similarity Search (HSS) [5], which drastically enhances characterization of compounds without exact matches in the library. This is a significant advantage for forensic applications such as seized drug analysis. For example, HSS successfully characterizes mass spectra of novel fentanyl analogs [6]. The HSS is available as part of the National Institute of Standards and Technology (NIST) Mass Spectrometry Data Center's MS Search Program [10].

This work investigates clustering and multidimensional scaling (MDS) [11,12] techniques with applications to HSS data. The analysis focuses on electron ionization (EI) mass spectra of potential illicit drugs,

* Corresponding author.
E-mail addresses: melinda.kleczynski@nist.gov (M.A. Kleczynski), arunmoorthy@trentu.ca (A.S. Moorthy), anthony.kearsley@nist.gov (A.J. Kearsley).

https://doi.org/10.1016/j.forc.2026.100795
Received 30 July 2026; Received in revised form 22 October 2026; Accepted 28 October 2026
Available online 6 November 2026
2468-1709/Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

but the techniques can also be applied to other types of compounds. We analyze collections of mass spectra consisting of a query spectrum together with a hit list. The hit list consists of all the mass spectra in a library whose similarity with the query spectrum is sufficiently high. HSS hit lists are known to include relatively large numbers of compounds, even for high Match Factor cutoffs [7], making visualization and clustering techniques helpful for interpretation. In addition, clustering is often performed using the output of a dimensionality reduction algorithm, rather than with the original mass spectral data [13]. Therefore, it is of interest to investigate the preservation of cluster structure by techniques such as MDS.

All elements of the hit list have high HSS Match Factors with the query spectrum, but some of these mass spectra have low HSS Match Factors with each other. This is a structural feature of the HSS, which is designed to emphasize the similarity of compounds which differ by a single modification. Note that by modification, we are generally referring to modifications which cause one or more shifted peaks in the mass spectrum, as opposed to modifications on neutral sites which are not detected by mass spectrometry. In contrast to HSS, approaches such as Transitive Alignment, developed for tandem mass spectrometry, are meant to compare compounds which differ by more than one modification [14].

In general, the mass spectra of compounds which differ by two modifications will not have high HSS Match Factors. If two compounds each differ from the query compound by one modification, then they may differ from each other by two modifications and thus their spectral comparison will likely yield a low HSS Match Factor. This can lead to interesting cluster structure in HSS hit lists.

One way to visualize this structure is through MDS. Given pairwise dissimilarity data, MDS produces coordinates whose pairwise distances approximate the original pairwise dissimilarities. This is typically achieved by minimizing a loss function in the stress [15] family; these are the most frequently used loss functions in modern MDS [11, 16,17]. The loss functions utilized in this work are STRESS (introduced as and often referred to as raw stress [15]) and SSTRESS [18]. These loss functions will be defined in Section 2.12. We follow the convention of some authors [18,19] of writing STRESS in all capital letters, although it is not intended here to be viewed as an acronym. The extra "S" in SSTRESS refers to the presence of squared dissimilarities and distances in the loss function [18].

MDS produces interpretable visualizations of mass spectral similarity and dissimilarity data, streamlining tasks such as detection of outliers and potential errors [20]. Principal components analysis (PCA) has also been used for applications including processing mass spectra of fentanyl analogs [21]. However, PCA does not accommodate the full range of dissimilarities which can be used with MDS.

The Mass Spectral Similarity Mapping algorithm, which utilizes non-metric MDS of HSS dissimilarities, is able to reveal groups of fentanyl analogs [22]; this algorithm is accessible as part of the NIST Fentanyl Classifier. *Non-metric MDS* generates configurations of points such that the *ordering of the distances* between MDS points approximates the *ordering of the dissimilarities* between data objects. In contrast, we consider *metric MDS* techniques which produce configurations such that the *distances* between MDS points approximate the *dissimilarities* between data objects. This facilitates the use of algorithms requiring numerical distances as inputs, such as standard clustering techniques.

The Mass Spectral Similarity Mapping algorithm successfully revealed unexpected cluster structure through the use of MDS [22]. Based on this finding, we aim to explore why clusters in HSS datasets are difficult to detect using the original dissimilarity data. Furthermore, we will suggest a modification to standard clustering algorithms so that clusters can be found directly from the original dissimilarities. We will also investigate whether the clusters found in MDS configurations are reflective of structure present in the original dataset. These inquiries are complicated by the availability of multiple MDS and clustering algorithms.

Finally, we will investigate some MDS implementation details. The primary issues we consider are the preservation of cluster structure, robustness with respect to the initial configuration used during optimization, and the effect of the optimization loss function. Analysis of three query spectra with different distributions of cluster sizes demonstrates a range of behavior. Based on the specific loss function that is used, the MDS configurations may have differing attributes, such as changes in the placement of the query spectrum with respect to other clusters in the dataset. This may affect subsequent interpretation of the similarity mapping.

2. Methods

The workflow presented in this manuscript is intended to give the analyst a better understanding of a query electron ionization (EI) mass spectrum and a hit list obtained using the Hybrid Similarity Search (HSS). Each dataset consists of the query spectrum and its HSS hit list, presented as a matrix whose elements are the HSS Match Factors for each pair of mass spectra. This matrix is symmetrized and then converted from a Symmetric HSS Match Factor matrix to an HSS dissimilarity matrix. These dissimilarities are the appropriate input format for subsequent clustering and MDS algorithms. In addition, the Simple Similarity Search (SSS) [23,24] Match Factors to the query spectrum are utilized in one step of the analysis, namely in the formation of a Query Cluster.

The HSS dissimilarity data is clustered by a two-step process. The first step is to form a Query Cluster by comparing the HSS and SSS Match Factors. The second step is to cluster the remaining spectra using standard clustering procedures.

This modified, two-step clustering procedure for the original HSS dissimilarity data forms the initial portion of the analysis. We also generate MDS configurations from the HSS dissimilarity data. We perform standard clustering for the MDS configurations, and compare the results.

2.1. Overview

The individual steps of the analysis are listed below. This process is largely assembled from standard data analysis techniques. The formation of the Query Cluster is the main deviation from typical pipelines.

1. Perform complete linkage and single linkage hierarchical clustering of the HSS dissimilarities. The results may be visualized using dendrograms and by sorting the rows and columns of the HSS dissimilarity matrix using the ordering of the dendrogram leaves.
2. Check Silhouette Coefficients for a range of numbers of clusters to verify that well-characterized cluster structure is not already apparent in the original HSS dissimilarities.
3. Compare SSS and Symmetric HSS Match Factors to the query spectrum and form the Query Cluster.
4. Cluster the HSS dissimilarities of the remaining spectra, using Silhouette Coefficients to choose the number of clusters.
5. Generate STRESS and SSTRESS MDS configurations for multiple random initial configurations. Select the final configurations with the lowest loss function values.
6. Perform complete and single linkage clustering of the MDS configurations.
7. Compare modified, two-step clustering results of the HSS dissimilarities to standard clustering results of the MDS configurations.

2.2. Programming language and packages

All analysis is performed using the Python programming language, version 3.13.2. The package Matplotlib [25], version 3.10.1, is used for plotting. NumPy [26], version 2.2.5, is used for numerical computa-

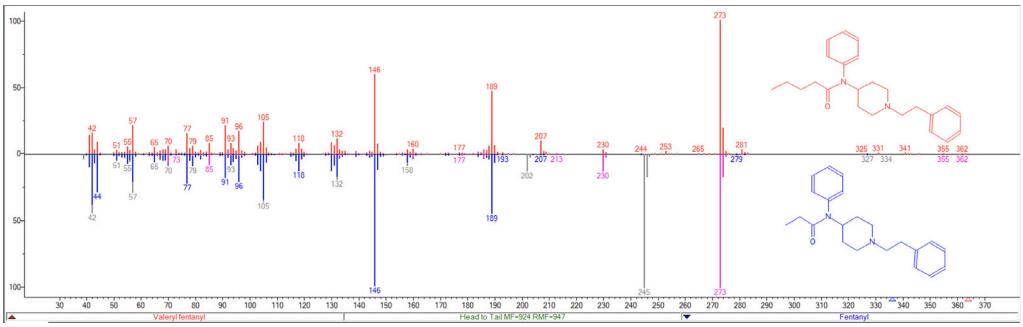


Fig. 1. A head-to-tail plot comparing the mass spectra of Valeryl fentanyl (top) and Fentanyl (bottom) as contained in the Scientific Working Group for the Analysis of Seized Drugs (SWGDRUG) [6] 3.14 library and visualized using NIST MS Search 3.0. The peaks contained in the original spectrum of Fentanyl are colored in blue and gray. The match factor computed between the spectrum of Valeryl fentanyl and the original spectrum of Fentanyl is 575. A hybrid spectrum generated by shifting peaks from the Fentanyl mass spectrum to optimize matches with Valeryl fentanyl can be constructed with the blue and pink peaks. The HSS Match Factor (computed using the spectrum of Valeryl fentanyl and the hybrid spectrum of Fentanyl optimized with Valeryl fentanyl) is 975.

tions. The pandas [27,28] package, version 2.2.3, is used for reading CSV files into DataFrames. SciPy [29], version 1.15.2, is used for optimization and clustering. We use scikit-learn [30] version 1.6.1 for quantifying clustering results. Standard techniques are the building blocks for the methods described in this manuscript, and consequently the analysis can also be implemented in other programming languages.

2.3. Hybrid similarity search (HSS)

For convenience, we provide a brief overview of the HSS algorithm; computational details are available elsewhere [5,31,32]. The key innovation of the HSS is the formation of a hybrid spectrum prior to similarity quantification via the Simple Similarity Search (SSS) [23,24]. The hybrid spectrum is constructed by performing peak shifting operations based on the difference of the nominal masses of the query and library compounds. An example is shown in Fig. 1. This improves the match in the event that two spectra differ by peaks that have been shifted by a single mass - often indicative of a single modification between the compounds. The HSS is especially advantageous in forensic applications, as the library may not contain an exact match to the query spectrum.

2.4. Symmetric match factors

Let $\mathrm{hMF}(i,j)$ denote the HSS Match Factor of mass spectra i and j. In general, HSS Match Factors are asymmetric. That is, it is possible that $\mathrm{hMF}(i,j) \neq \mathrm{hMF}(j,i)$. Averaging these values by setting

$$\mathrm{hMF}_S(i,j) = \frac{1}{2}\Big(\mathrm{hMF}(i,j) + \mathrm{hMF}(j,i)\Big) \tag{1}$$

yields $\mathrm{hMF}_S(i,j) = \mathrm{hMF}_S(j,i)$. The quantity in (1) is referred to as the Symmetric HSS Match Factor.

2.5. Datasets

The data are obtained from the NIST 23 electron ionization (EI) mass spectral library [33,34]. We analyze four datasets consisting of matrices of pairwise HSS Match Factors [5] of mass spectra. The four datasets were selected to demonstrate a range of sensitivity to changes in the MDS loss function. The first two datasets contain Fentanyl and analogs, while the third and fourth datasets contain amphetamine compounds. Each dataset is generated by selecting a query spectrum and including all available spectra whose Symmetric HSS Match Factors with the query spectrum are greater than or equal to the Symmetric HSS Match Factor cutoffs listed in Table 1. In other words, each dataset consists of a query spectrum and its hit list. The query spectra, the minimum Symmetric HSS Match Factors required for inclusion in the hit list, and the resulting numbers of mass spectra are given in Table 1. The NIST numbers and compound names of the mass spectra included in the four datasets can be found in the Supplementary Material. Throughout the manuscript, we refer to each dataset by giving the compound name for the query spectrum and, when needed to distinguish the datasets, the cutoff value.

We choose lower Symmetric HSS Match Factor cutoffs for the Valeryl fentanyl and 3,5-Dimethoxyamphetamine datasets due to the smaller numbers of mass spectra in these hit lists. Users can adjust these cutoffs as desired. In the extremes, very high cutoffs may exclude relevant clusters, while very low cutoffs may lead to poor cluster structure as questionable matches are included.

A unique structural feature of the HSS as compared to the SSS is that two mass spectra may have high Symmetric HSS Match Factors with the query spectrum, and very low Symmetric HSS Match Factors with each other. The datasets in this work were selected to have this property. Datasets can be screened for this behavior by computing the minimum Symmetric HSS Match Factor to the query spectrum and the minimum Symmetric HSS Match Factor in the entire dataset. The relevant values for the datasets considered in this work are shown in Table 2. This can also be seen visually in plots of the HSS dissimilarities, by comparing the query row/column to the rest of the matrix (as in Fig. 2).

In this work, we restricted attention to datasets consisting of a query spectrum and its hit list. Larger and more complex datasets may not be represented well by the techniques described in this manuscript. Larger hit lists with lower Symmetric HSS Match Factor cutoffs may have less clear cluster structure compared to the same query spectrum with a higher cutoff. We will explore this somewhat through the use of two different cutoff values for 3,5-Dimethoxyamphetamine. Other cases may show more extreme loss of structure as the cutoff is decreased.

2.6. Pairwise dissimilarities

Match Factors quantify similarity of mass spectra, but MDS operates on dissimilarities. As a data processing step, we convert from a similarity to a dissimilarity. HSS Match Factors and Symmetric HSS Match Factors range from a minimum possible value of 0 to a maximum possible value of 999. The pairwise dissimilarity δ_{ij} of mass spectra i and j can be defined as

$$\delta_{ij} = 1 - \frac{\mathrm{hMF}_S(i,j)}{999}. \tag{2}$$

The quantity in (2), which we refer to as the HSS dissimilarity, has values between 0 and 1, inclusive, with identical spectra having a dissimilarity of 0.

Table 1
Analysis is conducted for four datasets obtained from the NIST 23 EI mass spectral library [33,34]. Each dataset consists of a query spectrum and its hit list.

NIST number	Query compound	Drug type	Match cutoff	Number of spectra (n)
530558	Fentanyl	Fentanyl and analogs	900	61
463844	Valeryl fentanyl	Fentanyl and analogs	800	30
554261	3,5-Dimethoxyamphetamine	Amphetamines	800	34
554261	3,5-Dimethoxyamphetamine	Amphetamines	700	54

Table 2
Comparison of Symmetric HSS Match Factor (hMF_S) values in each dataset. Note the relatively low values in the final column.

Query compound	Match cutoff	Min. hMF_S to query	Min. hMF_S in dataset
Fentanyl	900	903	125
Valeryl fentanyl	800	869.5	455
3,5-Dimethoxyamphetamine	800	804	91
3,5-Dimethoxyamphetamine	700	700.5	28.5

2.7. Clustering

We use complete linkage clustering, which has been applied to HSS data [5], as well as single linkage clustering, which can have better performance when elongated clusters are present. These clustering methods are advantageous because they can be applied to non-metric dissimilarities, and do not require coordinate data. As agglomerative hierarchical clustering [35,36] methods, they have associated dendrograms which show the dissimilarities at which clusters are successively merged; this can inform the choice of number of clusters as well as revealing multiscale structure.

We begin by performing standard complete and single linkage clustering of the entire dataset. We also customize the clustering procedure for HSS hit lists by first assigning the query spectrum and other query-like spectra to a Query Cluster before proceeding with complete and single linkage clustering of the remaining mass spectra. Some compounds, such as many positional isomers, are known to generate very similar mass spectra [3,37,38]. Such spectra can be one source of query-like spectra in addition to the query spectrum.

The HSS algorithm is designed to increase similarity quantification for compounds separated by a single modification. We utilize this attribute to create a criterion for inclusion of spectra in the Query Cluster. We assign a spectrum to the Query Cluster if its SSS Match Factor with the query spectrum is equal to its Symmetric HSS Match Factor with the query spectrum. Note that the query spectrum will always qualify; both the SSS Match Factor and Symmetric HSS Match Factor of the query spectrum with itself are equal to 999.

Although this was effective for the datasets considered in this work, we suggest two possible extensions which may be helpful in other datasets. First, it is possible that shifting small peaks when constructing the hybrid spectrum may slightly increase the HSS Match Factor for compounds that are not separated by a modification detectable in the mass spectra. If clustering still performs poorly after removing the Query Cluster, mass spectra with similar but not identical SSS and Symmetric HSS Match Factors with the query spectrum could be considered for possible inclusion in the Query Cluster. Second, if compounds that differ from the query compound by more than one modification are included in the dataset, these could also generate identical SSS and Symmetric HSS Match Factors with the query spectrum. This could occur, for example, when using lower cutoff values to define the hit list. In this case, it may be desirable to separate and then cluster the spectra having identical SSS and Symmetric HSS Match Factors with the query spectrum, instead of placing them all in a single Query Cluster.

2.8. Cluster evaluation

One of the tools we use for cluster evaluation and choosing the number of clusters is the Silhouette Coefficient [39]. This can be computed using standard software packages; we use the implementation in scikit-learn [30]. The algorithm is also summarized here for interested readers. The Silhouette Coefficient can be computed from a list of cluster labels and either a pairwise dissimilarity matrix or a collection of point coordinates.

Given an object i and clusters $C_1, \ldots, C_k$, let $C(i)$ be the cluster containing object i. The quantity $\rho(i, C_j)$ denotes the average dissimilarity from object i to the objects in cluster C_j, excluding the dissimilarity from i to itself in the event that $C_j = C(i)$. If i is the only object in cluster $C(i)$, then set $s(i) = 0$. Otherwise, one computes

$$\begin{aligned} a(i) &= \rho(i, C(i)), \\ b(i) &= \min_{j=1,\ldots,k, C_j \neq C(i)} \rho(i, C_j), \\ s(i) &= \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}. \end{aligned} \tag{3}$$

The Silhouette Coefficient of object i is given by $s(i)$ in (3).

The maximum possible value of $s(i)$ is one, and values close to one indicate a good cluster assignment for object i. Values close to zero indicate that object i could be moved to some other cluster without substantially changing the quality of its cluster assignment. Negative values, particularly those close to negative one, indicate that object i may have been assigned to the wrong cluster.

In this work, prospective numbers of clusters range from two through ten. The final number of clusters is chosen to be the one which produces the highest mean Silhouette Coefficient. One advantage of the Silhouette Coefficient is that values are assigned to individual data objects, enabling additional evaluation on a spectrum by spectrum basis.

In addition to computing Silhouette Coefficients, various visual inspections aid in verifying that distinct clusters are present and that the selected number of clusters is appropriate. One such approach is to plot the dendrogram associated with the complete linkage clustering. Ideally, there should be a wide range of HSS dissimilarity thresholds such that cutting the dendrogram at that threshold will yield the given number of clusters. In addition, sorting the rows and columns of the HSS dissimilarity matrices according to the order of the dendrogram leaves facilitates visual verification of the similarity of rows/columns corresponding to mass spectra which cluster together.

2.9. Comparison of clustering results

Clustering the points of an MDS configuration may produce different results than clustering the original HSS dissimilarities. One method for quantifying the similarity of two clustering results, which is informative for the datasets considered in this work, is the Fowlkes-Mallows Index (FMI) [40]. The FMI is a general technique for clustering comparison, but we describe it here in the particular case of comparing a clustering result obtained from HSS dissimilarities and a clustering result obtained from an MDS configuration.

Given the quantities

- A: Number of pairs of spectra which are in the same cluster in both the HSS dissimilarity clustering and the MDS clustering,

- B: Number of pairs of spectra which are in the same cluster in the HSS dissimilarity clustering but different clusters in the MDS clustering,
- C: Number of pairs of spectra which are in different clusters in the HSS dissimilarity clustering but the same cluster in the MDS clustering,

the FMI is defined as $A/\sqrt{(A+B)(A+C)}$. We use the Fowlkes-Mallows implementation from scikit-learn [30]. Small values of the FMI (smallest possible value is zero) indicate poor agreement of the HSS dissimilarity clustering and the MDS clustering, while larger values (largest possible value is one) indicate good agreement.

2.10. MDS initial configurations

MDS algorithms based on iterative optimization require an initial configuration of points. In general, MDS is sensitive to the initial configuration used [19,41–43]. One common strategy is to generate multiple random initial configurations, and choose the final configuration which has the lowest loss function value. A procedure for generating random initial configurations [19] is summarized here. Throughout, n is the number of objects and p is the MDS dimension. In the current work, objects are mass spectra and $p = 2$. The process works for higher values of p if desired. The initial configuration of points (with no points constrained to lie on the origin or one of the axes) is stored in an $n \times p$ matrix, denoted X^0. For $p = 2$, X^0 has the form

$$X^0 = \begin{pmatrix} x_1 & y_1 \\ x_2 & y_2 \\ x_3 & y_3 \\ \vdots & \vdots \\ x_n & y_n \end{pmatrix}. \tag{4}$$

For larger values of p, the initial configuration matrix in (4) would be modified by adding additional columns.

The scale of the random initial configuration can be chosen based on the scale of the dissimilarities of the objects to be represented. This is achieved by defining

$$\sigma = \left(\frac{\sum_{i<j} \delta_{ij}^2}{pn(n-1)} \right)^{1/2}. \tag{5}$$

Then the random initial configuration is given by $X^0 = \sigma Z$, where σ is defined as in (5) and Z is an $n \times p$ matrix whose elements are pseudorandom numbers from the standard normal distribution. The factor σ is defined so that the expected squared distance between points in the initial configuration is the same as the mean squared dissimilarity between data objects.

2.11. Fixing MDS translation and rotation

To facilitate comparing MDS configurations, one can standardize the translation, rotation, and reflection of the configurations. This does not change the clustering results, but eases visual inspection and coordinate comparisons. One method for accomplishing this [19] is summarized here for the case when $p = 2$. The idea is to constrain one point of the configuration to lie at the origin and another point to lie on the x-axis. This can be implemented by setting some coordinates equal to zero, producing the configuration matrix

$$X^k = \begin{pmatrix} 0 & 0 \\ x_2 & 0 \\ x_3 & y_3 \\ \vdots & \vdots \\ x_n & y_n \end{pmatrix} \tag{6}$$

where k denotes the optimization iteration.

In the case of $k = 0$, in other words the initial configuration, the matrix (6) is obtained from (4) by setting the values of x_1, y_1, and y_2 equal to zero. These values remain equal to zero in every following configuration throughout the iterative optimization process. This fixes the translation and rotation of the final MDS configurations, but it does not resolve potential reflection over one or both of the axes. This will be addressed through a post-processing alignment step.

In this work, the query spectrum is chosen as the object whose MDS point is fixed at the origin. This amounts to a choice of translation of the MDS configurations. This selection has the benefit of making the location of the query spectrum readily apparent in plots of the MDS configurations.

One strategy for choosing an object whose MDS point must lie on the x-axis is to make a selection that encourages a large range of values of x. This mimics the behavior of PCA, for which the first component has the largest explained variance ratio. In some datasets, one could choose an object with a high dissimilarity with the object whose MDS point is at the origin. However, for HSS dissimilarities and objects consisting of the query spectrum and its hit list, every object has a small pairwise dissimilarity with the object whose MDS point is at the origin (the query spectrum). Instead, our selection is based on the columns of the matrix of HSS Match Factors. The mass spectrum whose column is the most different from the column of the query spectrum (where the column difference is defined as the sum of the absolute values of the differences of the column entries) is selected to have its MDS point constrained to the x-axis.

2.12. MDS optimization

The goal of MDS is to find an optimal configuration of points, whose coordinates form the matrix X, such that the Euclidean ("straight line") distances between MDS points approximate the original (potentially non-Euclidean) pairwise dissimilarities from the dataset. Technically, distances other than the Euclidean distance can be used for points in the MDS configuration [15], but this approach is atypical and not considered in this work. A common approach is to use iterative optimization to minimize one of a family of loss functions of the form

$$L_r(X) = \sum_{i<j} \left(\left(\delta_{ij}^2\right)^r - \left(d_{ij}^2(X)\right)^r \right)^2. \tag{7}$$

Here, $d_{ij}(X)$ is the Euclidean distance between points i and j in the MDS configuration stored in the matrix X. The quantity δ_{ij} is the pairwise dissimilarity between data objects i and j, in this case defined in (2).

Setting $r = \frac{1}{2}$ in the general loss function (7) yields the STRESS loss function, and setting $r = 1$ gives the SSTRESS loss function. For both values of r, this can be approached as a nonlinear least-squares problem. These can be solved by a range of techniques; the current work uses the Levenberg–Marquardt algorithm [44] accessed through SciPy [29]. We compute both STRESS and SSTRESS configurations for the four datasets. Each plotted configuration is the one with the minimum loss function value out of 1000 random initial configurations. This is larger than the number of random initial configurations typically used if reproducibility analysis is not of interest, and the only objective is to obtain one final MDS configuration for subsequent visualization or clustering.

The residuals are given by $r_{ij} = \delta_{ij} - d_{ij}(X)$. There is an interpretable formulation of the STRESS and SSTRESS loss functions defined in (7) based on the residuals [45]. For STRESS,

$$L_{1/2}(X) = \sum_{i<j} r_{ij}^2. \tag{8}$$

For SSTRESS,

$$L_1 = \sum_{i<j} \left(\delta_{ij} + d_{ij}(X)\right)^2 r_{ij}^2. \tag{9}$$

In other words, Eqs. (8) and (9) mean that the STRESS loss function is the sum of the squares of the residuals, while in the SSTRESS loss function, residuals associated with large dissimilarities and large MDS distances have a greater contribution.

In addition to least-squares optimization using the Levenberg–Marquardt algorithm, we also produced STRESS MDS configurations using the scikit-learn [30] implementation of SMACOF (Scaling by Majorizing a Complicated Function) [11,46] for optimization. For the STRESS MDS configurations with the lowest loss function values, least-squares and SMACOF optimization produced consistent clustering results, as demonstrated by Fowlkes-Mallows Indices being equal to 1. More detailed results are shown in the Supplementary Material. The main text focuses on optimization via least-squares, as this is applicable to both STRESS and SSTRESS loss functions.

2.13. Fixing MDS reflection

The final MDS configurations have standardized translation and rotation due to constraining a selected point to the origin and a second selected point to the x-axis. However, distinct MDS configurations could differ by reflection over one or both of the axes. This can be corrected by multiplying the x-coordinates and/or y-coordinates of each configuration by -1 to ensure compatible layouts.

The point which is constrained to the x-axis is assigned to the positive x-axis by, if necessary, multiplying the x-coordinates of individual MDS configurations by -1. A similar procedure can be followed for the y-coordinates once a designated point is chosen to restrict to having a positive y-coordinate (assuming we can always find a point that does not lie on the x-axis). In this analysis, the point with the highest y-coordinate in the minimum STRESS configuration is selected. These point selections suffice for the current work. In general, there may exist points which are poor indicators of the need to perform reflections, such as those lying near the origin.

3. Theory

This section contains a brief explanation of why MDS representations cannot replicate all the patterns in HSS data when hit lists contain diverse compounds. The following discussion applies to MDS in general, and not only our specific implementation. Suppose mass spectra A and B both have very high HSS Match Factors with query spectrum Q, but very low HSS Match Factors with each other. This could occur, for example, if the compounds corresponding to mass spectra A and B differ from the query compound by one modification, and differ from each other by two modifications. HSS Match Factors are converted to dissimilarities to perform MDS. High HSS Match Factors correspond to low dissimilarities, and vice versa. Thus, the conversion yields dissimilarities δ_{AQ} and δ_{BQ} which are much smaller than δ_{AB}; in other words, $\delta_{AQ} \ll \delta_{AB}$ and $\delta_{BQ} \ll \delta_{AB}$. For δ_{AQ} and δ_{BQ} both sufficiently small, their sum is still much smaller than δ_{AB}:

$$\delta_{AQ} + \delta_{BQ} \ll \delta_{AB}. \tag{10}$$

One can use some form of either metric or non-metric MDS to produce a configuration of points corresponding to these mass spectra. The dissimilarities between mass spectra, namely δ_{AQ}, δ_{BQ}, and δ_{AB}, correspond to Euclidean ("straight line") distances d_{AQ}, d_{BQ}, and d_{AB} between points in the MDS configuration. Euclidean distances must satisfy a property known as the triangle inequality; for these three distances, the triangle inequality stipulates that

$$d_{AB} \leq d_{AQ} + d_{BQ}. \tag{11}$$

Comparing the inequalities (10) and (11) shows that the relationships between some dissimilarities and sums of dissimilarities will be distorted when producing any Euclidean MDS configuration of a sufficiently diverse hit list characterized by HSS. This is a key motivation for comparing different methods for producing MDS configurations, as the particular distortions observed may depend on the technique used.

4. Results and discussion

4.1. Hierarchical clustering of HSS dissimilarities

Step 1 of the analysis consists of examining the standard hierarchical clustering structure of the original HSS dissimilarity matrix. Complete and single linkage clustering are hierarchical clustering techniques. Here, we focus on visualization of the complete linkage structure. The complete linkage dendrograms (before separating the Query Cluster) are shown in the left column of Fig. 2. The ordering of the dendrogram leaves is a convenient way to sort the rows and columns of each HSS dissimilarity matrix. This has the benefit of being independent of the chosen number of clusters, in the event of multiscale structure being present. For Fentanyl and its hit list, shown in Fig. 2(a), similar rows/columns are generally grouped together by this sorting. However, there is a single row/column (labeled "Q") of all low HSS dissimilarities (plotted in dark purple) which is mixed in with another group of similar appearing rows/columns. This row/column corresponds to the query compound, Fentanyl. The query spectrum has low HSS dissimilarities to every spectrum in the hit list. The query spectrum is placed securely in one of the clusters by the complete linkage clustering algorithm, as illustrated in the dendrogram. The query spectrum may not be distinguishable from the appearance of the dendrogram, but it is visible in the sorted HSS dissimilarity matrix due to its low maximum HSS dissimilarity to any other spectrum. For Valeryl fentanyl and 3,5-Dimethoxyamphetamine, there are multiple query-like spectra which have low HSS dissimilarities to all other spectra in the dataset. Complete linkage clustering can group these with various clusters.

We performed the two-step clustering procedure using complete and single linkage clustering of the original HSS dissimilarities for all datasets. We found that, for the best number of clusters identified, the complete linkage results were the same as the single linkage results. Therefore, we do not distinguish between the complete and single linkage results for the original HSS dissimilarities. The NIST numbers and compound names for the mass spectra assigned to each cluster are shown in the Supplementary Material.

4.2. Silhouette coefficients before removal of query cluster

If the original HSS dissimilarities can be readily clustered by standard methods, then removal of the Query Cluster is not necessary. In Step 2 of the analysis, the quality of clustering in the original datasets is assessed by viewing the individual Silhouette Coefficients for a range of numbers of clusters. The results are shown in the left column of Fig. 3. For all datasets, all numbers of clusters evaluated, and both clustering methods, there are spectra whose Silhouette Coefficients are negative or near zero. Therefore, we conclude that there are spectra which do not have good cluster assignments, and proceed with removal of the Query Cluster.

4.3. Comparison of SSS and symmetric HSS match factors

Step 3 of the analysis consists of selecting spectra for inclusion in the Query Cluster. This is done by comparing the SSS and Symmetric HSS Match Factors to the query spectrum. These are plotted in Fig. 4. The points lying on the dashed line represent mass spectra whose SSS and Symmetric HSS Match Factors to the query spectrum are equal; these spectra form the Query Cluster. The query spectrum itself has coordinates (999, 999) in this plot, and is included in the Query Cluster.

4.4. Silhouette coefficients after removal of query cluster

In Step 4 of the analysis, we again compute and plot the Silhouette Coefficients of the original HSS dissimilarities, but this time excluding

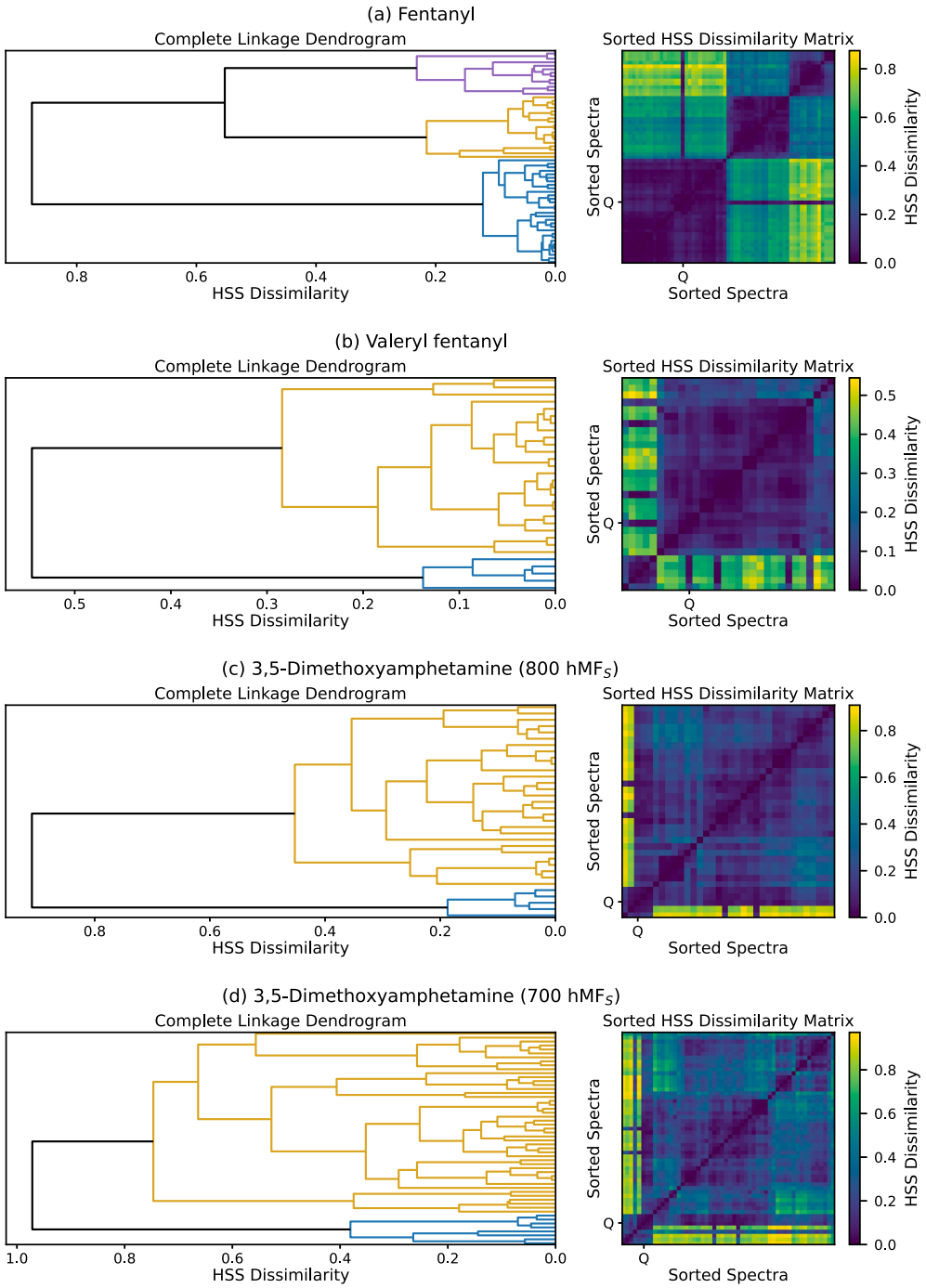


Fig. 2. The complete linkage clustering dendrograms for the (a) Fentanyl, (b) Valeryl fentanyl, (c) 3,5-Dimethoxyamphetamine (800 hMF$_S$ cutoff), and (d) 3,5-Dimethoxyamphetamine (700 hMF$_S$ cutoff) datasets are shown in the left column. The corresponding HSS dissimilarity matrices, with the rows and columns sorted by the dendrogram leaves, are shown in the right column. The label Q indicates the query spectrum.

the Query Cluster. The results are shown in the right column of Fig. 3. For the Fentanyl dataset, the best number of clusters is 3 for both complete and single linkage clustering. There are no points with excessively low individual Silhouette Coefficients. For the Valeryl fentanyl dataset, the best number of clusters is 2, and again no points have markedly low Silhouette Coefficients. Note that the Query Cluster is in addition to the best number of clusters determined by the Silhouette Coefficients, so the number of clusters in the MDS configurations will be four for Fentanyl and three for Valeryl fentanyl.

For the two 3,5-Dimethoxyamphetamine datasets, prior to removing the Query Cluster, there were individual spectra with negative Silhouette Coefficients for every tested number of clusters and for both complete and single linkages. After removing the Query Cluster, the Silhouette Coefficients were best for two clusters. For the 800 hMF$_S$ cutoff, all individual Silhouette Coefficients were greater than 0.65. However, for the 900 hMF$_S$ cutoff, there is a single mass spectrum with a low Silhouette Coefficient (0.214) for both complete and single linkage clustering. The corresponding compound is N-acetyl-3,4-dimethoxyamphetamine.

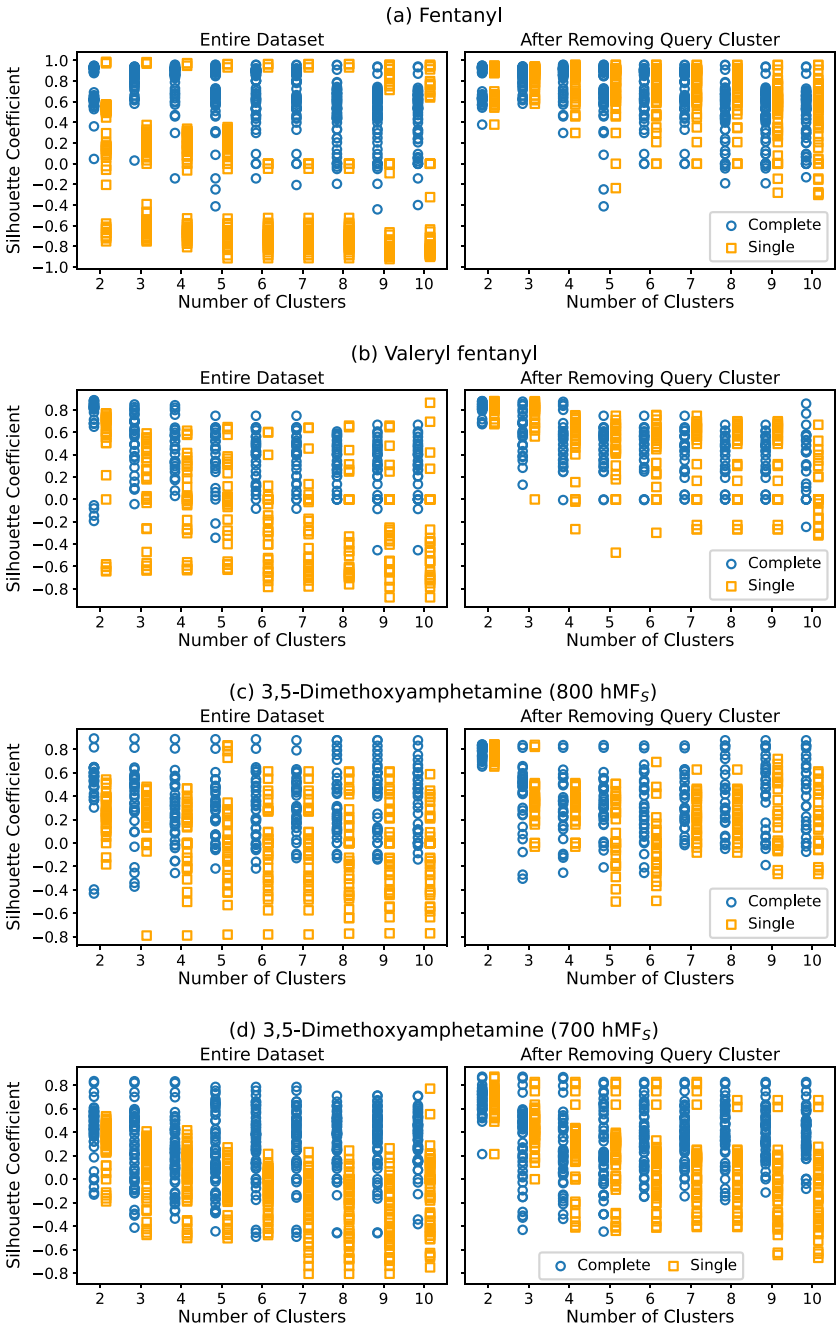


Fig. 3. Silhouette Coefficients for individual spectra in the (a) Fentanyl, (b) Valeryl fentanyl, (c) 3,5-Dimethoxyamphetamine (800 hMF$_S$ cutoff), and (d) 3,5-Dimethoxyamphetamine (700 hMF$_S$ cutoff) datasets. Silhouette Coefficients close to one are indicative of good clustering performance, while values which are close to zero, or even negative, indicate questionable results.

4.5. MDS configurations and clustering

Step 5 consists of computing MDS configurations from multiple random initial configurations. The final configurations with the lowest values of the STRESS and SSTRESS loss functions are shown in Figs. 5, 6, 7, and 8. The STRESS configurations are in the left column of each figure, and the SSTRESS configurations are in the right column.

Step 6 consists of clustering the MDS configurations. The top row of Figs. 5, 6, 7, and 8 shows the complete linkage clustering results, while the bottom row shows the single linkage results. These cluster labels are computed from the MDS configurations using standard clustering (no special step is required to remove the Query Cluster). The number of clusters is taken to be one more than the best number of clusters indicated by the Silhouette Coefficients in the right column of Fig. 3.

The results for the Fentanyl dataset, as shown in Fig. 5, appear similar for both MDS loss functions and both clustering methods. However, differences begin to appear for the other datasets. We see that the SSTRESS configurations may have a small, condensed cluster around the query spectrum at the origin. In addition, some configurations contain an elongated group of points which is separated by complete but not single linkage clustering.

When computing the MDS configurations, we constrained one point to lie at the origin and another point to lie on the x-axis, chosen to ideally produce a large range of values of x. As shown in Figs. 5, 6, 7,

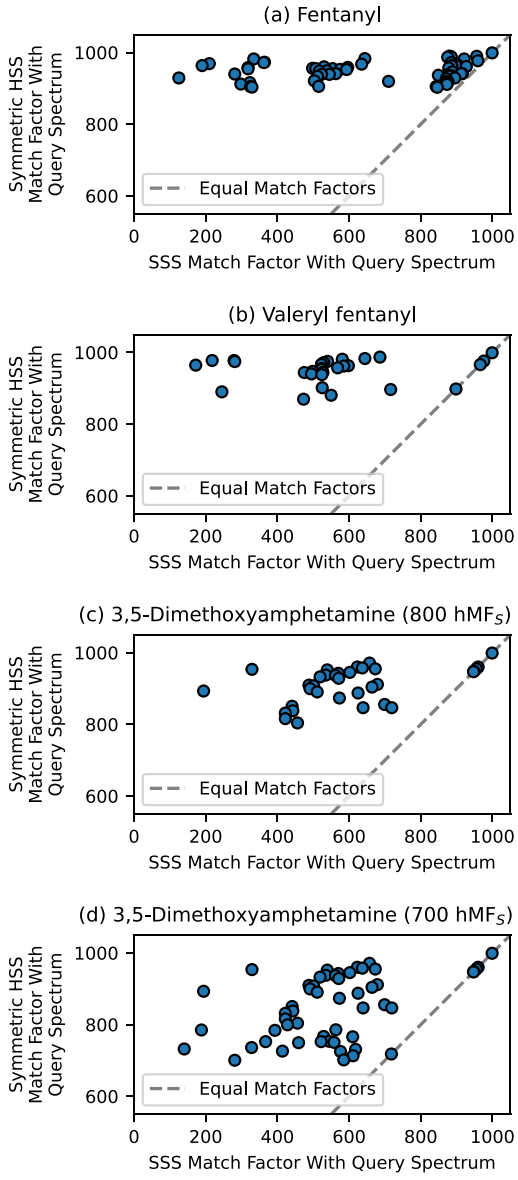


Fig. 4. Comparison of Simple Similarity Search (SSS) and Symmetric Hybrid Similarity Search (HSS) Match Factors for the (a) Fentanyl, (b) Valeryl fentanyl, (c) 3,5-Dimethoxyamphetamine (800 hMF$_S$ cutoff), and (d) 3,5-Dimethoxyamphetamine (700 hMF$_S$ cutoff) datasets. Each point lying on the dashed line represents a mass spectrum whose SSS Match Factor with the query spectrum is equal to its Symmetric HSS Match Factor with the query spectrum.

and 8, this seems to have been effective. In many cases, the clusters are separable using only the first component (MDS 1).

4.6. Comparison of HSS dissimilarity clustering and MDS clustering

The analysis concludes with Step 7, comparison of the clustering results. The results for the MDS configurations with the lowest loss function values are listed in Table 3. Complete agreement of the clustering results yields a Fowlkes-Mallows Index equal to 1.

For the Fentanyl dataset, all clustering results from the MDS configurations are consistent with the results from the two-step clustering procedure of the original HSS dissimilarities. This is independent of the MDS loss function (STRESS versus SSTRESS) and clustering method (complete versus single linkage). The solitary placement of Fentanyl in the MDS configurations is consistent with it being the only spectrum assigned to the Query Cluster in the two-step procedure for the HSS dissimilarities.

There is more variation in the Fowlkes-Mallows Indices for the other datasets. The best value (closest to 1) occurs for single linkage clustering of the SSTRESS MDS configuration. The 3,5-Dimethoxyamphetamine (700 hMF$_S$ cutoff) dataset is the only one for which a Fowlkes-Mallows Index of 1 is never attained. This is due to a disagreement in cluster assignment for the single spectrum which had a low Silhouette Coefficient in the original HSS dissimilarity data.

In addition to the MDS configurations with the lowest loss function values, we computed Fowlkes-Mallows Indices for all MDS configurations, in order to explore robustness of MDS clustering results with respect to the initial configuration used in optimization. As before, the Fowlkes-Mallows Index compares the MDS clustering results to the two-step HSS dissimilarity clustering results. The results are shown in Figs. 9, 10, 11, and 12.

In practice, when using random initial configurations for performing MDS, it is common to try multiple random initial configurations and use the final MDS configuration with the lowest loss function value. This protects against occurrences such as choosing an unfortunate initial configuration and terminating at a local minimum which could have easily been avoided with more optimization starting points. When viewing the Fowlkes-Mallows Indices, we separate the results into two groups, less than the median of all loss function values (blue histograms), and greater than or equal to the median (orange histograms). Consistency is desirable, as is seen with the Fentanyl results in Fig. 9. The exception is some STRESS configurations clustered via complete linkage, but these have high STRESS values and would likely be filtered out when using multiple random initial configurations. On the other hand, for the 3,5-Dimethoxyamphetamine (700 hMF$_S$ cutoff) dataset shown in Fig. 12, consistency is better for the SSTRESS configurations than the STRESS configurations.

We found that SSTRESS can be a favorable MDS loss function in terms of clustering performance, as demonstrated in Table 3. However, some MDS implementations do not produce SSTRESS configurations [19]. Fortunately, framing the SSTRESS problem as a nonlinear least-squares problem allows MDS to be performed using standard optimization packages, making the techniques described in this manuscript accessible to a wide range of users.

4.7. Discussion

For the datasets considered in this work, we were able to find clustering and MDS methods such that the cluster structure of 2-dimensional MDS configurations was similar to the cluster structure of the original HSS dissimilarities. This was demonstrated by the high Fowlkes-Mallows Indices in Table 3. This may not be the case for all datasets. In particular, representation in low-dimensional spaces may be challenging for datasets with large numbers of clusters. If the dendrogram and Silhouette Coefficients of the original dataset indicate a strong cluster structure, but this is not preserved in 2-dimensional MDS configurations, higher dimensions could be trialed. Using the lowest dimension possible produces MDS configurations which are more suitable for visualization. HSS dissimilarities do not behave like Euclidean distances, so not all structure is preserved in MDS configurations, even for high dimensions.

MDS configurations may have qualitatively different structural features for different choices of MDS loss function. In this work, we focused on the difference between STRESS and SSTRESS configurations. Analysts may be interested in exploring other variants of these loss functions.

A major advantage of the HSS is its ability to characterize compounds which do not have an exact match in a library, a possible outcome in forensic applications. Another technique which has

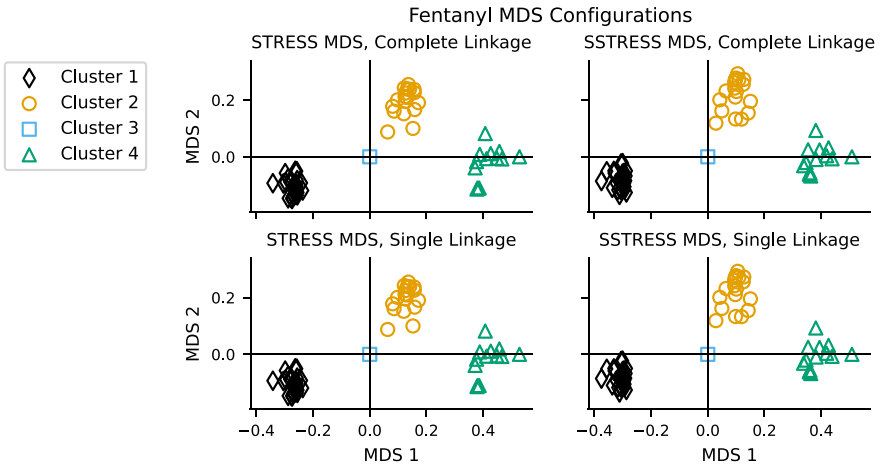


Fig. 5. The STRESS and SSTRESS MDS configurations with the lowest values of the loss functions for the Fentanyl dataset. Each point in a configuration represents a mass spectrum, with the point at the origin representing the query spectrum.

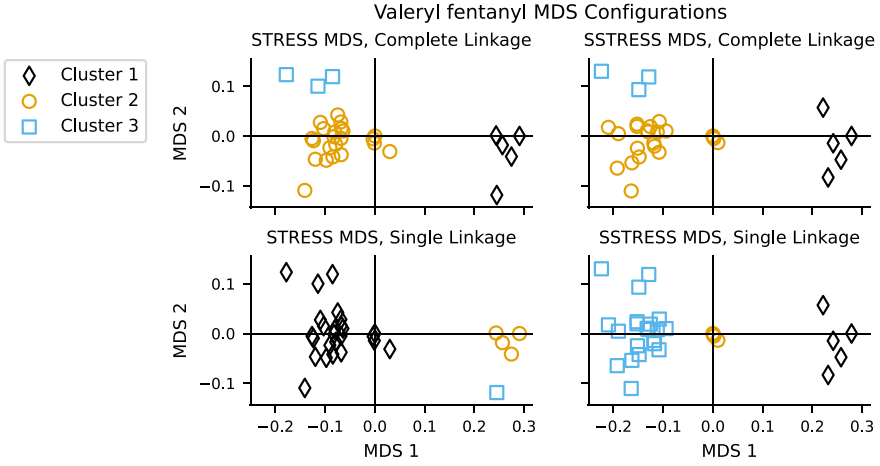


Fig. 6. The STRESS and SSTRESS MDS configurations with the lowest values of the loss functions for the Valeryl fentanyl dataset. Each point in a configuration represents a mass spectrum, with the point at the origin representing the query spectrum.

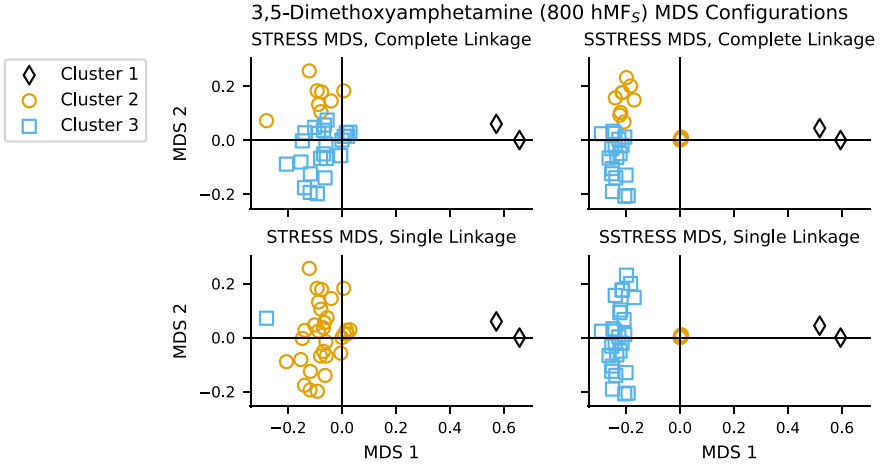


Fig. 7. The STRESS and SSTRESS MDS configurations with the lowest values of the loss functions for the 3,5-Dimethoxyamphetamine (800 hMF$_S$) dataset. Each point in a configuration represents a mass spectrum, with the point at the origin representing the query spectrum.

been utilized to handle novel substances is molecular networking; this has been employed, for example, for the analysis of cosine similarities between tandem mass spectra of potential illicit drugs [47]. There are interesting possibilities for combining the methods described in this manuscript with molecular networking. For example, MDS is a well-established technique for choosing node locations for network drawing [48–50]. Networks constructed using HSS Match Factors rather than cosine similarities may benefit from using SSTRESS MDS, as discussed in this work, to choose the node locations for visualization.

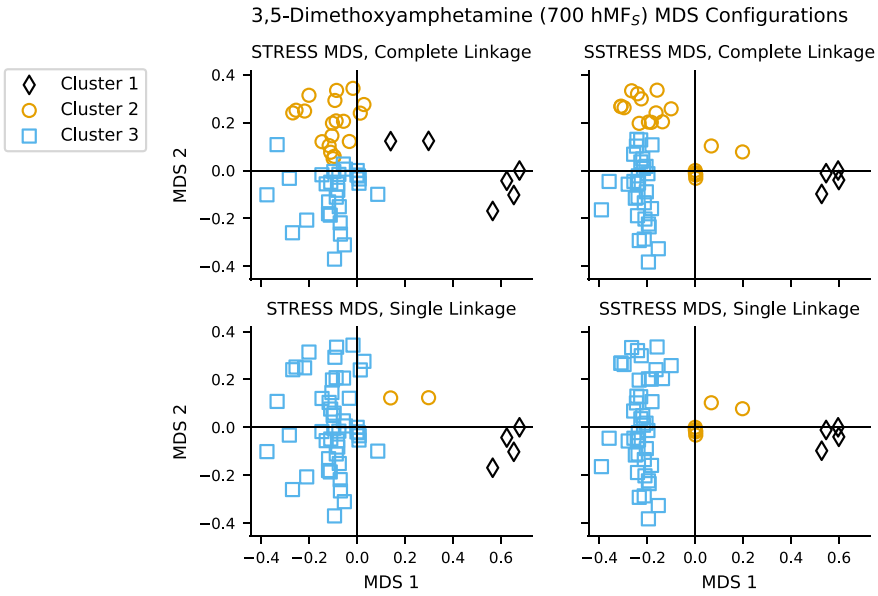


Fig. 8. The STRESS and SSTRESS MDS configurations with the lowest values of the loss functions for the 3,5-Dimethoxyamphetamine (700 hMF$_S$) dataset. Each point in a configuration represents a mass spectrum, with the point at the origin representing the query spectrum.

Table 3
Fowlkes-Mallows Indices for clustering of the original dissimilarities versus the MDS configurations with the lowest values of the loss functions.

Query compound	Match cutoff	Clustering method	STRESS MDS	SSTRESS MDS
Fentanyl	900 hMF$_S$	Complete linkage	1.00	1.00
		Single linkage	1.00	1.00
Valeryl fentanyl	800 hMF$_S$	Complete linkage	0.732	0.732
		Single linkage	0.844	1.00
3,5-Dimethoxyamphetamine	800 hMF$_S$	Complete linkage	0.632	0.672
		Single linkage	0.818	1.00
3,5-Dimethoxyamphetamine	700 hMF$_S$	Complete linkage	0.612	0.700
		Single linkage	0.877	0.974

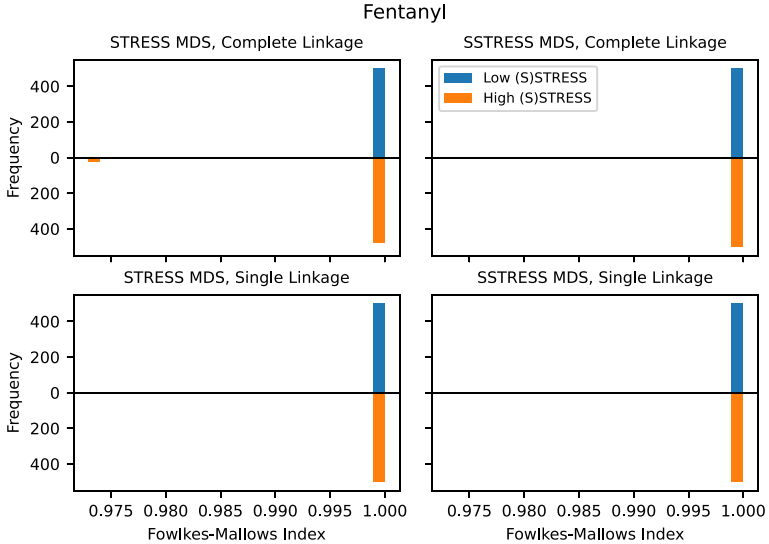


Fig. 9. Distribution of Fowlkes-Mallows Indices (FMIs) for MDS configurations of the Fentanyl dataset. Comparisons are made between cluster labels from the original HSS dissimilarities and cluster labels from the MDS configurations. FMI values of 1 indicate perfect agreement of the clustering results.

One limitation of the current work is that we did not consider how variations in the HSS Match Factors could affect the resulting MDS configurations. Each dataset consisted of a fixed matrix of Match Factors. Various aspects of the spectral comparison, such as the location of the modification [51] and the mass spectral deconvolution method [52], can affect HSS efficacy. It would be interesting to explore these considerations in future work. We are also interested in exploring MDS of replicate measurements of fentanyl analogs in future work,

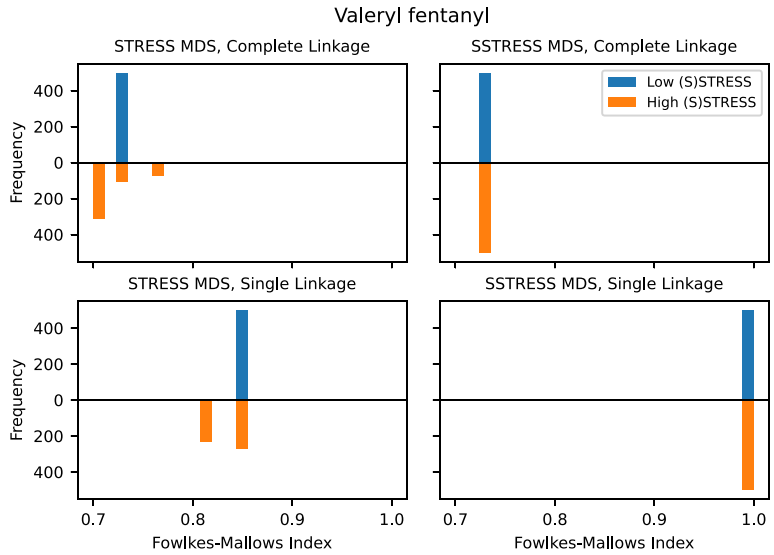


Fig. 10. Distribution of Fowlkes-Mallows Indices (FMIs) for MDS configurations of the Valeryl fentanyl dataset. Comparisons are made between cluster labels from the original HSS dissimilarities and cluster labels from the MDS configurations. FMI values of 1 indicate perfect agreement of the clustering results.

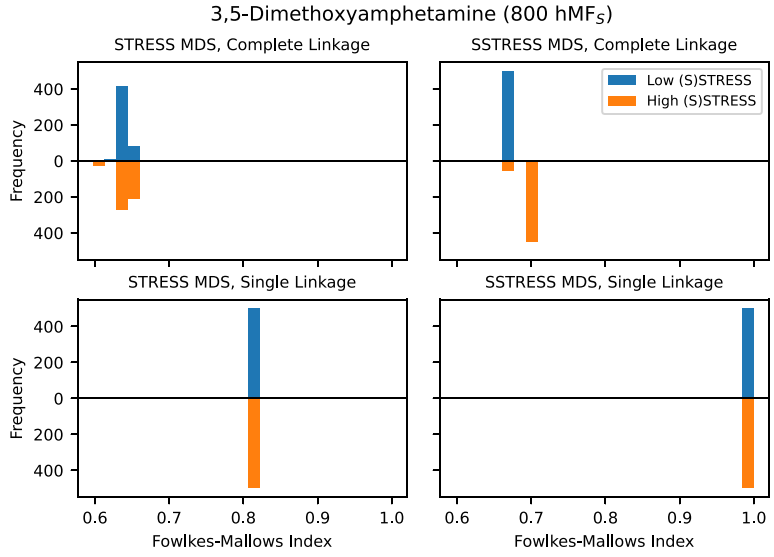


Fig. 11. Distribution of Fowlkes-Mallows Indices (FMIs) for MDS configurations of the 3,5-Dimethoxyamphetamine (800 hMF$_S$ cutoff) dataset. Comparisons are made between cluster labels from the original HSS dissimilarities and cluster labels from the MDS configurations. FMI values of 1 indicate perfect agreement of the clustering results.

as has been done with PCA of Surface-Enhanced Raman Spectroscopy (SERS) [53].

There is interest in developing new ways to quantify similarity and dissimilarity in order to improve the performance of forensic algorithms; for example, this has been discussed in the context of direct analysis in real time mass spectrometry (DART-MS) [54]. If new, complex notions of similarity and dissimilarity continue to be introduced, then visualization and compatibility with other algorithms will become increasingly challenging. MDS will be a valuable tool for preprocessing and dimensionality reduction. However, this leads to new questions about the structure preserved by MDS configurations. Adjustments may be needed, such as the changes to standard clustering pipelines employed in this work.

Hybrid search implementations are available for EI [5] and tandem [31,55] mass spectral data. The techniques described in this manuscript were applied to mass spectral datasets from the NIST 23 EI library [33,34]. Hybrid search for tandem mass spectra has produced interesting results for applications including human milk oligosaccharides [56], erectile dysfunction drugs [57], and controlled substances [58]. Notably, the controlled substance application involved combining HSS Match Factors with artificial intelligence tools as part of an "identifier layer" for analyzing tandem mass spectral data [58]. Future work could explore the application of our methods to tandem mass spectral datasets.

5. Conclusions

We described clustering and multidimensional scaling procedures with general applicability to Hybrid Similarity Search (HSS) results, demonstrated for hit lists of Fentanyl, Valeryl fentanyl, and 3,5-Dimethoxyamphetamine. These datasets have interesting structure with respect to complete and single linkage clustering. The complete linkage dendrograms suggest a strong signature of a small number of clusters. Sorting the rows and columns of the HSS dissimilarity matrices according to the order of the dendrogram leaves placed similar rows and columns together, with the exception of rows/columns corresponding to the query spectrum (for all query spectra considered) and other query-like spectra (for Valeryl fentanyl

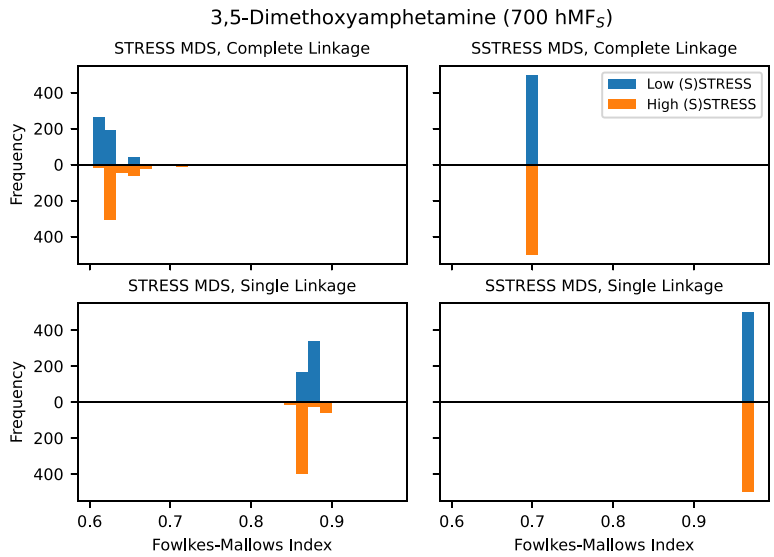


Fig. 12. Distribution of Fowlkes-Mallows Indices (FMIs) for MDS configurations of the 3,5-Dimethoxyamphetamine (700 hMF$_S$ cutoff) dataset. Comparisons are made between cluster labels from the original HSS dissimilarities and cluster labels from the MDS configurations. FMI values of 1 indicate perfect agreement of the clustering results.

and 3,5-Dimethoxyamphetamine). We found that inclusion of the query spectrum and other query-like spectra is associated with clustering results in which some data points have low individual Silhouette Coefficients. It is not unexpected that these spectra would perform poorly with respect to this or other quantifications of clustering quality, because they are close to all elements of the hit list, including those placed in other clusters.

We described a two-step clustering procedure for HSS dissimilarities which involves assigning the query spectrum and other query-like spectra to a Query Cluster before performing standard clustering of the remaining spectra. For the datasets considered, this improves the overall Silhouette Coefficients. Furthermore, we were able to find MDS and clustering methods such that the two-step clustering of the HSS dissimilarities and the standard clustering of the MDS configurations were the same or very similar, indicating that some MDS configurations excel at preserving this cluster structure.

Clustering is a widespread technique in forensic data analysis, and reproducibility is a major concern [59]. In some cases, MDS configurations are relatively robust to the choice of loss function, while in other situations STRESS and SSTRESS MDS produce substantially different results. In these cases, the best agreement with the two-step HSS dissimilarity clustering results was produced by single linkage clustering of the SSTRESS MDS configurations. We also investigated reproducibility of the least-squares MDS configurations with respect to differing random initial configurations, and found that for some datasets, SSTRESS MDS was more robust to the choice of optimization starting point.

The HSS is an invaluable tool for the description of forensic data, and other mass spectral datasets for which a query spectrum may not have an exact match in the library. Care should be used when combining HSS dissimilarities with other algorithms, since these dissimilarities do not behave like Euclidean distances. In particular, two potentially advantageous adjustments to standard pipelines are separate treatment of the query spectrum and other query-like spectra in clustering algorithms, and the use of SSTRESS rather than STRESS for performing MDS.

Conversion of complex dissimilarity or similarity data to low-dimensional Euclidean features is an analysis step that can be seen in a range of forensic workflows. For example, recent work for the Catalan Regional Police utilizing near-infrared spectroscopy data successfully used Kendall correlations and custom weighted spectral differences to produce four features for subsequent analysis based on Euclidean distances [60]. MDS is a general method designed to generate low-dimensional configurations from dissimilarity or similarity inputs, facilitating further analysis and visualization. The specific challenges encountered in this work, and the steps taken to address them, may serve as a starting point for incorporating MDS in other forensic analyses.

CRediT authorship contribution statement

Melinda A. Kleczynski: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Formal analysis, Data curation, Conceptualization. **Arun S. Moorthy:** Writing – review & editing, Methodology, Data curation, Conceptualization. **Anthony J. Kearsley:** Writing – review & editing, Supervision, Methodology, Conceptualization.

Disclaimer

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

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.

Appendix A. Supplementary data

Tables of the mass spectra in each cluster for the four datasets analyzed, figures of SMACOF MDS configurations, and a table comparing least-squares and SMACOF clustering results.

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.forc.2026.100795.

Data availability

Data and code can be found at https://github.com/usnistgov/cluster-mass-spec-hybrid-search.

References

[1] M. Joshi, E. Sisco, Forensic seized drug analysis: Current challenges and emerging analytical solutions, WIREs Forensic Sci. 5 (5) (2023) e1486, http://dx.doi.org/10.1002/wfs2.1486.

[2] C.A. Valdez, Gas chromatography-mass spectrometry analysis of synthetic opioids belonging to the fentanyl class: A review, Crit. Rev. Anal. Chem. 52 (8) (2022) 1938–1968, http://dx.doi.org/10.1080/10408347.2021.1927668.

[3] M.A. Bodnar 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, http://dx.doi.org/10.1016/j.forsciint.2016.11.013.

[4] S. Stein, Mass spectral reference libraries: An ever-expanding resource for chemical identification, Anal. Chem. 84 (2012) 7274–7282, http://dx.doi.org/10.1021/ac301205z.

[5] A.S. Moorthy, W.E. Wallace, A.J. Kearsley, D.V. Tchekhovskoi, S.E. Stein, Combining fragment-ion and neutral-loss matching during mass spectral library searching: A new general purpose algorithm applicable to illicit drug identification, Anal. Chem. 89 (2017) 13261–13268, http://dx.doi.org/10.1021/acs.analchem.7b03320.

[6] W. Ji, W.E. Wallace, Comprehensive data evaluation methods used in developing the SWGDRUG mass spectral reference library for seized drug identification, Anal. Chem. 96 (2024) 17004–17012, http://dx.doi.org/10.1021/acs.analchem.4c04425.

[7] D.F. McGlynn, L.D. Yee, H.M. Garraffo, L.Y. Geer, T.D. Mak, Y.A. Mirokhin, D.V. Tchekhovskoi, C.N. Jen, A.H. Goldstein, A.J. Kearsley, S.E. Stein, New library-based methods for nontargeted compound identification by GC-EI-MS, J. Am. Soc. Mass Spectrom. 36 (2025) 389–399, http://dx.doi.org/10.1021/jasms.4c00451.

[8] O.D. Sparkman, Evaluating electron ionization mass spectral library search results, J. Am. Soc. Mass Spectrom. 7 (1996) 313–318, http://dx.doi.org/10.1016/1044-0305(95)00705-9.

[9] A.S. Moorthy, A.J. Kearsley, Pattern similarity measures applied to mass spectra, in: M. Cruz, C. Parés, P. Quintela (Eds.), Progress in Industrial Mathematics: Success Stories, Springer International Publishing, Cham, 2021, pp. 43–53, http://dx.doi.org/10.1007/978-3-030-61844-5_4.

[10] 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 (5) (2023) 1484–1493, http://dx.doi.org/10.1111/1556-4029.15284.

[11] I. Borg, P.J.F. Groenen, Modern Multidimensional Scaling: Theory and Applications, second ed., Springer, New York, NY, 2005, http://dx.doi.org/10.1007/0-387-28981-X.

[12] I. Borg, P.J.F. Groenen, P. Mair, Applied Multidimensional Scaling and Unfolding, second ed., Springer, Cham, Switzerland, 2018, http://dx.doi.org/10.1007/978-3-319-73471-2.

[13] A.G. Beck, M. Muhoberac, C.E. Randolph, C.H. Beveridge, P.R. Wijewardhane, H.I. Kenttämaa, G. Chopra, Recent developments in machine learning for mass spectrometry, ACS Meas. Sci. Au 4 (3) (2024) 233–246, http://dx.doi.org/10.1021/acsmeasuresciau.3c00060.

[14] X. Wang, M. Strobel, A.T. Aron, V.V. Phelan, D.D. Acharya, C.J. Brown, K. Clevenger, J. Hu, A. Kretsch, E.H. Mahood, C. Menegatti, Q. Xiong, M. Wang, Network topology evaluation and transitive alignments for molecular networking, J. Am. Soc. Mass Spectrom. 35 (9) (2024) 2165–2175, http://dx.doi.org/10.1021/jasms.4c00208.

[15] J.B. Kruskal, Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis, Psychometrika 29 (1964) 1–27, http://dx.doi.org/10.1007/BF02289565.

[16] P.J.F. Groenen, M. van de Velden, Multidimensional scaling by majorization: A review, J. Stat. Softw. 73 (8) (2016) 1–26, http://dx.doi.org/10.18637/jss.v073.i08.

[17] T. Rusch, P. Mair, K. Hornik, Cluster optimized proximity scaling, J. Comput. Graph. Stat. 30 (4) (2021) 1156–1167, http://dx.doi.org/10.1080/10618600.2020.1869027.

[18] Y. Takane, F.W. Young, J. de Leeuw, Nonmetric individual differences multidimensional scaling: An alternating least squares method with optimal scaling features, Psychometrika 42 (1977) 7–67, http://dx.doi.org/10.1007/BF02293745.

[19] A.J. Kearsley, R.A. Tapia, M.W. Trosset, The solution of the metric STRESS and SSTRESS problems in multidimensional scaling using Newton's method, Comput. Statist. 13 (1998) 369–396.

[20] W.E. Wallace, W. Ji, D.V. Tchekhovskoi, K.W. Phinney, S.E. Stein, Mass spectral library quality assurance by inter-library comparison, J. Am. Soc. Mass Spectrom. 28 (4) (2017) 733–738, http://dx.doi.org/10.1007/s13361-016-1589-4.

[21] N. Gilbert, R.E. Mewis, O.B. Sutcliffe, Classification of fentanyl analogues through principal component analysis (PCA) and hierarchical clustering of GC–MS data, Forensic Chem. 21 (2020) 100287, http://dx.doi.org/10.1016/j.forc.2020.100287.

[22] A.S. Moorthy, A.J. Kearsley, W.G. Mallard, W.E. Wallace, Mass spectral similarity mapping applied to fentanyl analogs, Forensic Chem. 19 (2020) 100237, http://dx.doi.org/10.1016/j.forc.2020.100237.

[23] S.E. Stein, Estimating probabilities of correct identification from results of mass spectral library searches, J. Am. Soc. Mass Spectrom. 5 (4) (1994) 316–323, http://dx.doi.org/10.1016/1044-0305(94)85022-4.

[24] S.E. Stein, D.R. Scott, Optimization and testing of mass spectral library search algorithms for compound identification, J. Am. Soc. Mass Spectrom. 5 (9) (1994) 859–866, http://dx.doi.org/10.1016/1044-0305(94)87009-8.

[25] J.D. Hunter, Matplotlib: A 2D graphics environment, Comput. Sci. Eng. 9 (3) (2007) 90–95, http://dx.doi.org/10.1109/MCSE.2007.55.

[26] C.R. Harris, K.J. Millman, S.J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N.J. Smith, R. Kern, M. Picus, S. Hoyer, M.H. van Kerkwijk, M. Brett, A. Haldane, J.F. del Río, M. Wiebe, P. Peterson, P. Gérard-Marchant, K. Sheppard, T. Reddy, W. Weckesser, H. Abbasi, C. Gohlke, T.E. Oliphant, Array programming with NumPy, Nature 585 (7825) (2020) 357–362, http://dx.doi.org/10.1038/s41586-020-2649-2.

[27] W. McKinney, Data structures for statistical computing in Python, in: Stéfan van der Walt, Jarrod Millman (Eds.), Proceedings of the 9th Python in Science Conference, 2010, pp. 56–61, http://dx.doi.org/10.25080/Majora-92bf1922-00a.

[28] The pandas development team, pandas-dev/pandas: Pandas (v2.2.3), 2024, http://dx.doi.org/10.5281/zenodo.13819579.

[29] P. Virtanen, R. Gommers, T.E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S.J. van der Walt, M. Brett, J. Wilson, K.J. Millman, N. Mayorov, A.R.J. Nelson, E. Jones, R. Kern, E. Larson, C.J. Carey, İ. Polat, Y. Feng, E.W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E.A. Quintero, C.R. Harris, A.M. Archibald, A.H. Ribeiro, F. Pedregosa, P. van Mulbregt, SciPy 1.0 Contributors, SciPy 1.0: Fundamental algorithms for scientific computing in Python, Nat. Methods 17 (2020) 261–272, http://dx.doi.org/10.1038/s41592-019-0686-2.

[30] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine learning in Python, J. Mach. Learn. Res. 12 (2011) 2825–2830, URL http://jmlr.org/papers/v12/pedregosa11a.html.

[31] M.C. Burke, Y.A. Mirokhin, D.V. Tchekhovskoi, S.P. Markey, J. Heidbrink Thompson, C. Larkin, S.E. Stein, The hybrid search: A mass spectral library search method for discovery of modifications in proteomics, J. Proteome Res. 16 (5) (2017) 1924–1935, http://dx.doi.org/10.1021/acs.jproteome.6b00988.

[32] B.T. Cooper, X. Yan, Y. Simón-Manso, D.V. Tchekhovskoi, Y.A. Mirokhin, S.E. Stein, Hybrid search: A method for identifying metabolites absent from tandem mass spectrometry libraries, Anal. Chem. 91 (21) (2019) 13924–13932, http://dx.doi.org/10.1021/acs.analchem.9b03415.

[33] National Institute of Standards and Technology (NIST), Mass Spectrometry Data Center, 2026, https://chemdata.nist.gov. (Accessed 26 April 2026).

[34] National Institute of Standards and Technology (NIST), NIST23: Updates to the NIST tandem and electron ionization spectral libraries, 2026, https://www.nist.gov/programs-projects/nist23-updates-nist-tandem-and-electron-ionization-spectral-libraries. (Accessed 26 April 2026).

[35] F. Murtagh, P. Contreras, Algorithms for hierarchical clustering: An overview, WIREs Data Min. Knowl. Discov. 2 (1) (2012) 86–97, http://dx.doi.org/10.1002/widm.53.

[36] F. Murtagh, P. Contreras, Algorithms for hierarchical clustering: An overview, II, WIREs Data Min. Knowl. Discov. 7 (6) (2017) e1219, http://dx.doi.org/10.1002/widm.1219.

[37] M. Verhoeven, J. Bonetti, R. Kranenburg, A. van Asten, Chemical identification and differentiation of positional isomers of novel psychoactive substances – A comprehensive review, Trends Anal Chem. 166 (2023) 117157, http://dx.doi.org/10.1016/j.trac.2023.117157.

[38] A.S. Moorthy, E.P. Erisman, A.J. Kearsley, Y. Liang, E. Sisco, W.E. Wallace, On the challenge of unambiguous identification of fentanyl analogs: Exploring measurement diversity using standard reference mass spectral libraries, J. Forensic Sci. 68 (5) (2023) 1494–1503, http://dx.doi.org/10.1111/1556-4029.15322.

[39] P.J. Rousseeuw, Silhouettes: A graphical aid to the interpretation and validation of cluster analysis, J. Comput. Appl. Math. 20 (1987) 53–65, http://dx.doi.org/10.1016/0377-0427(87)90125-7.

[40] E.B. Fowlkes, C.L. Mallows, A method for comparing two hierarchical clusterings, J. Amer. Statist. Assoc. 78 (383) (1983) 553–569, http://dx.doi.org/10.1080/01621459.1983.10478008.

[41] I. Borg, P. Mair, The choice of initial configurations in multidimensional scaling: Local minima, fit, and interpretability, Austrian J. Stat. 46 (2) (2017) 19–32, http://dx.doi.org/10.17713/ajs.v46i2.561.

[42] M. Kleczynski, A. Kearsley, Topological initialization for multidimensional scaling, 2025, http://dx.doi.org/10.6028/NIST.TN.2349, Technical Note (NIST TN), National Institute of Standards and Technology, Gaithersburg, MD.

[43] M.A. Kleczynski, A.J. Kearsley, Clustering persistence barcodes of multidimensional scaling representations from random initial configurations, Austrian J. Stat. 55 (3) (2026) 91–112.
[44] J.J. Moré, The Levenberg-Marquardt algorithm: Implementation and theory, in: G.A. Watson (Ed.), Numerical Analysis, Springer, Berlin, Heidelberg, 1978, pp. 105–116, http://dx.doi.org/10.1007/BFb0067700.
[45] W.J. Heiser, Multidimensional scaling with least absolute residuals, in: H.H. Bock (Ed.), Classification and Related Methods of Data Analysis, 1988, pp. 455–462.
[46] J. de Leeuw, P. Mair, Multidimensional scaling using majorization: SMACOF in R, J. Stat. Softw. 31 (3) (2009) 1–30, http://dx.doi.org/10.18637/jss.v031.i03.
[47] J.S. Yu, H. Seo, G.B. Kim, J. Hong, H.H. Yoo, MS-based molecular networking of designer drugs as an approach for the detection of unknown derivatives for forensic and doping applications: A case of NBOMe derivatives, Anal. Chem. 91 (9) (2019) 5483–5488, http://dx.doi.org/10.1021/acs.analchem.9b00294.
[48] J.B. Kruskal, J.B. Seery, Designing network diagrams, in: Proceedings of the First General Conference on Social Graphics, 1980, pp. 22–50.
[49] T. Kamada, S. Kawai, An algorithm for drawing general undirected graphs, Inform. Process. Lett. 31 (1) (1989) 7–15, http://dx.doi.org/10.1016/0020-0190(89)90102-6.
[50] E.R. Gansner, Y. Koren, S. North, Graph drawing by stress majorization, in: J. Pach (Ed.), Graph Drawing, Springer, Berlin, Heidelberg, 2005, pp. 239–250, http://dx.doi.org/10.1007/978-3-540-31843-9_25.
[51] A. Pollard, J.T. Davidson, Investigating the effect of substitution location on fentanyl analog identification for methyl-substituted fentanyl analogs using GC-EI-MS, Forensic Chem. 36 (2023) 100534, http://dx.doi.org/10.1016/j.forc.2023.100534.
[52] C.K. Chua, Y. Lv, W. Zhao, Y. Ren, H.J. Zhang, Improving annotation of known-unknowns with accurately reconstructed mass spectra, Int. J. Mass Spectrom. 451 (2020) 116321, http://dx.doi.org/10.1016/j.ijms.2020.116321.
[53] S. Dogruer Erkok, E. Hernandez, J. Cruz, A.M. Mebel, B. McCord, Differentiating structurally similar fentanyl analogs by comparing density functional theory (DFT) calculations and surface-enhanced Raman spectroscopy (SERS) results, Appl. Spectrosc. 78 (7) (2024) 667–679, http://dx.doi.org/10.1177/00037028241246010.
[54] A.S. Moorthy, E. Sisco, A new library-search algorithm for mixture analysis using DART-MS, J. Am. Soc. Mass Spectrom. 32 (7) (2021) 1725–1734, http://dx.doi.org/10.1021/jasms.1c00097.
[55] M.C. Burke, Z. Zhang, Y.A. Mirokhin, D.V. Tchekovskoi, Y. Liang, S.E. Stein, False discovery rate estimation for hybrid mass spectral library search identifications in bottom-up proteomics, J. Proteome Res. 18 (9) (2019) 3223–3234, http://dx.doi.org/10.1021/acs.jproteome.8b00863.
[56] C.A. Remoroza, T.D. Mak, M.L.A. De Leoz, Y.A. Mirokhin, S.E. Stein, Creating a mass spectral reference library for oligosaccharides in human milk, Anal. Chem. 90 (15) (2018) 8977–8988, http://dx.doi.org/10.1021/acs.analchem.8b01176.
[57] I. Jang, J. Lee, J. Lee, B.H. Kim, B. Moon, J. Hong, H.B. Oh, LC–MS/MS software for screening unknown erectile dysfunction drugs and analogues: Artificial neural network classification, peak-count scoring, simple similarity search, and hybrid similarity search algorithms, Anal. Chem. 91 (2019) 9119–9128, http://dx.doi.org/10.1021/acs.analchem.9b01643.
[58] S.Y. Lee, S.T. Lee, S. Suh, B.J. Ko, H.B. Oh, Revealing unknown controlled substances and new psychoactive substances using high-resolution LC–MS-MS machine learning models and the hybrid similarity search algorithm, J. Anal. Toxicol. 46 (7) (2022) 732–742, http://dx.doi.org/10.1093/jat/bkab098.
[59] M. Bovens, B. Ahrens, I. Alberink, A. Nordgaard, T. Salonen, S. Huhtala, Chemometrics in forensic chemistry — Part I: Implications to the forensic workflow, Forensic Sci. Int. 301 (2019) 82–90, http://dx.doi.org/10.1016/j.forsciint.2019.05.030.
[60] R. Sáez Hernández, S. Soriano Hernández, M. Mazarío-García, R. López Glaría, M.A. Fiol Planellas, M. Alcalà Bernàrdez, Å. Rinnan, J. Cruz, A novel chemometric local approach for qualitative and quantitative analysis of cocaine, MDMA, and THC-related products: Method application within the Mossos d'Esquadra (Catalan Regional Police), Anal. Chem. 98 (12) (2026) 9250–9259, http://dx.doi.org/10.1021/acs.analchem.5c07967.