ACS Partner Journal

Journal of the American Society for Mass Spectrometry

pubs.acs.org/jasms

Research Article

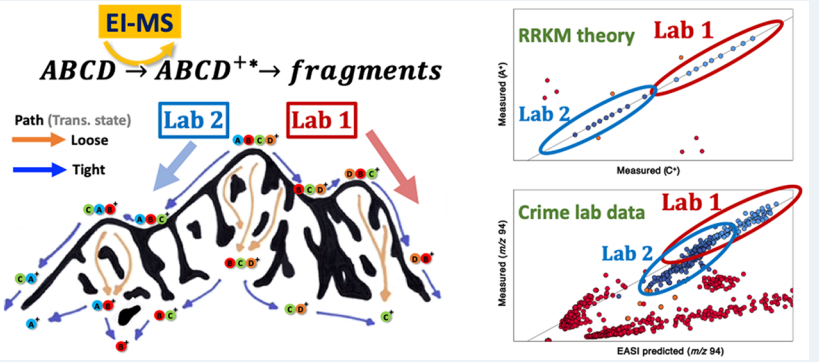
Expert Algorithm for Substance Identification Using Mass Spectrometry: Statistical Foundations in Unimolecular Reaction Rate Theory

Glen P. Jackson,* Samantha A. Mehnert, J. Tyler Davidson, Brandon D. Lowe, Emily A. Ruiz, and Jacob R. King

Cite This: *J. Am. Soc. Mass Spectrom.* 2023, 34, 1248−1262

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information



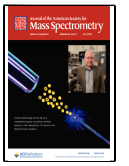
ABSTRACT: This study aims to resolve one of the longest-standing problems in mass spectrometry, which is how to accurately identify an organic substance from its mass spectrum when a spectrum of the suspected substance has not been analyzed contemporaneously on the same instrument. Part one of this two-part report describes how Rice−Ramsperger−Kassel−Marcus (RRKM) theory predicts that many branching ratios in replicate electron−ionization mass spectra will provide approximately linear correlations when analysis conditions change within or between instruments. Here, proof-of-concept general linear modeling is based on the 20 most abundant fragments in a database of 128 training spectra of cocaine collected over 6 months in an operational crime laboratory. The statistical validity of the approach is confirmed through both analysis of variance (ANOVA) of the regression models and assessment of the distributions of the residuals of the models. General linear modeling models typically explain more than 90% of the variance in normalized abundances. When the linear models from the training set are applied to 175 additional known positive cocaine spectra from more than 20 different laboratories, the linear models enabled ion abundances to be predicted with an accuracy of <2% relative to the base peak, even though the measured abundances vary by more than 30%. The same models were also applied to 716 known negative spectra, including the diastereomers of cocaine: allococaine, pseudococaine, and pseudoallococaine, and the residual errors were larger for the known negatives than for known positives. The second part of the manuscript describes how general linear regression modeling can serve as the basis for binary classification and reliable identification of cocaine from its diastereomers and all other known negatives.

KEYWORDS: *spectral comparisons, spectral algorithm, search algorithm, forensic science, compound identification, binary classification, drug identification*

INTRODUCTION

Since its first demonstration in the late 1930s,[1] electron ionization (née impact) mass spectrometry (EI-MS) has proven to be a powerful tool for identifying organic substances.[2−4] More than 80 years since its introduction, EI-MS in the form of gas chromatography−mass spectrometry (GC-MS) continues to be one of the most commonly employed methods of compound identification in forensic science,[5,6] metabolomics,[7,8] flavor and fragrance,[5] toxicology,[9] and pharmacology.[10] Given that there are now more than 300,000 compounds in the latest EI-MS database from NIST, computerized approaches to compound identification have long been a necessity.[11−16]

Received: March 16, 2023
Revised: May 12, 2023
Accepted: May 15, 2023
Published: May 31, 2023



Ignoring the importance of GC retention times for a moment, the most common approach to identifying a substance from its mass spectrum is by performing a library search, using either the apex spectrum of a GC peak or the average spectrum across a GC peak. Library search algorithms then compare the presence and abundance of peaks in the questioned mass spectrum (the unknown) to those in known exemplar spectra of reference samples in a database. Depending on the database vendor and the algorithm employed, the software typically returns a list of the top 10 closest matches with some measure of spectral similarity assigned to each pairwise comparison between the questioned spectrum and the reference spectra.[17,18]

Once a best candidate is chosen, an analyst, such as a seized drug analyst, is usually required to manually compare the unknown and reference spectra. Organizations such as the World Anti-Doping Agency (WADA), United Nations Office on Drugs and Crime (UNODC), and the American Society for Testing and Materials (ASTM) all provide acceptance criteria that analysts are expected to use when evaluating the peak abundances of a questioned sample relative to those of a reference spectrum.[19−25] These acceptance criteria generally permit the abundances of peaks in the questioned spectrum to fall within ±20% of their respective abundances in the reference spectrum,[19−25] although agencies vary in their stringencies.[26,27] These acceptance criteria are generally in agreement with typical uncertainties observed in replicate spectra of standards collected in crime laboratories.[28]

The use of spectral libraries is complicated by the fact that spectra derive from various types of mass analyzers, and the spectra vary in their quality, the presence of background ions, and the occurrence of mass bias.[29] Furthermore, instruments have different geometries, operating conditions, and other idiosyncrasies that contribute to the overall variance in ion abundances of replicate spectra.[15,28,30] The choice of tuning algorithm and tuning frequency also plays a role in the variability of replicate spectra.[31] These differences within and between instruments have deleteriously affected the ability of algorithms to rank or identify substances correctly.[12,13,15,17,18,32−36] All of these factors negatively affect the ability of current algorithms to make true positive identifications, even when a reference spectrum of the questioned substance is present in the database. For example, most algorithms typically only provide around 80% accuracy in ranking the correct identity in the #1 position.[13,37−44] For reasons that will become apparent later, the inclusion of replicate spectra of each substance can significantly improve identification rates.[12,45,46]

Various methods have been developed to improve the confidence in substance identification using spectral comparison techniques, such as changing the weighting factors,[40,41,43,44,47] changing the peak selection or abundance-normalization method,[37,48] modifying the results based on experimental information after the spectrum has already been collected,[49] and increasing the size of the library.[12] Other improvements include the use of partial and semipartial correlations[50] and wavelet and Fourier transformations to increase the accuracy of the identification algorithms.[51]

Given the breadth of potential applications and the varying weight of false positives and false negatives in different applications, there is no consensus as to which approach, or algorithm, is "best". Therefore, analysts must select an approach that is simply the best fit for their purpose.[10,12,18] As indicated earlier, the most common approach to improving the success rate of mass spectral identifications is to combine the mass spectral information with independent information, such as the retention time or retention index.[7,8,40,49,52−59] However, if the unknown material has not or cannot be analyzed on the same instrument, the database retention times/indices may not be sufficiently reliable to enable the differentiation of structurally similar compounds. In cases involving coelution, chromatographic peaks can be deconvoluted from each other and/or from background ions and thereby increase the success of mass spectral identifications.[49,50,55,60−62]

THE RANDOM AND NONRANDOM VARIANCE OF REPLICATE SPECTRA

One aspect of spectral comparison algorithms that is often overlooked is the assumption, and oftentimes mathematical requirement, that any variance in the *relative* abundance of peaks within a replicate spectrum is randomly or independently variable at each m/z value.[48,63] Such a mathematical requirement has been assumed since the first use of computational approaches to background-subtraction[64] or spectral deconvolution into discrete component spectra,[60,65,66] whether using simultaneous linear equations[65] or matrix theory.[67,68] By default, deconvolution algorithms explicitly assume unit correlation among the *absolute* signals of fragments as a function of time, or scan number, and they implicitly assume that any unexplained variance in a given scan at a specific m/z value is random.[49,50,55,60,61,65,67−70] Furthermore, to have statistical validity, most measures of spectral similarity and dissimilarity between questioned and reference spectra also require independent variance, i.e., no correlation, in the *relative* abundance at each m/z value within replicate spectra.[15,37,71] As an example of this reliance, a recent and extremely effective approach to spectral comparisons uses combined unequal variance t-tests at each m/z value to compare questioned and known spectra. Combining the results of independent t-tests explicitly requires independent variability of each t-test to enable the computation of random match probabilities.[72−74] However, as indicated elsewhere,[28] and as we will show below, replicate spectra still contain strong correlations in the normalized abundances of peaks, so the different m/z values are not independently variable. Supplemental Figure S1 supports this contention by showing obvious correlations and anticorrelations in the normalized abundances of certain fragments of cocaine across a chromatographic peak.

THE VALUE OF CROSS-CORRELATIONS IN REPLICATE SPECTRA

In the 1960s, Crawford and Morrison noted that systematic differences in fragmentation patterns occur when a mass spectrum of a substance was collected by sweeping the magnetic field instead of the ion acceleration voltage of a magnetic sector instrument.[75] In 1979, van Marlen et al. also noted that the abundances of peaks in replicate spectra were not independently variable and that when spectra deviate from a reference spectrum, it was "virtually impossible to take the correlations between the errors for the different m/e (sic) values into account".[48] Here, an "error" is the deviate abundance at one m/z value between a questioned peak abundance and a reference peak abundance.

The difficulty in accounting for correlations between residuals has caused almost all search algorithms since then to assume that peak abundances in replicate spectra are independently variable,[76] including the traditional and popular peptide-scoring

https://doi.org/10.1021/jasms.3c00089
J. Am. Soc. Mass Spectrom. 2023, 34, 1248−1262

algorithms based on tandem mass spectra of protonated precursors.[36,77−81] However, by considering correlations between theoretical and measured fragment ions in tandem mass spectra of peptides, Fu et al. were able to reduce the peptide mismatch rate from their tandem mass spectra by as much as 10%.[82] Of course, such approaches are only possible when the sequence is known for a precursor ion and the theoretical fragment ions can be accurately predicted. Driver et al. also used covariance analysis between fragments in replicate spectra to improve the selectivity of their peptide identification algorithm, and they also demonstrated the ability to identify mechanistic relationships between fragment-fragment pairs, such as complementary b/y fragments or consecutive fragments, like the loss of CO from a b ion to form a corresponding a ion. Related to the present study, their covariance mapping was conducted on simple replicate spectra that were collected without deliberate variance in the magnitude or mechanism of perturbation.[83−86] Zhang's group showed that partial and semipartial correlations could reduce the false discovery rate of library searches of EI spectra by a few percent.[50] Therefore, the ability to incorporate cross-correlations of replicate spectra into mass spectral comparison algorithms is a demonstrated mechanism to improve their performance.

Mass bias and spectral tilting can be thought of as m/z-dependent correlations in which low mass ions correlate with one another, high mass ions correlate with one another, but low mass and high mass ions anticorrelate with one another. McLafferty's group and Dromey have both shown that by incorporating correction terms, or scaling terms, to account for mass bias in replicate spectra, the unexplained variance in replicate peak abundances can be reduced by as much as 40%, and the reliability of search algorithms can be increased by as much as 10%.[11,87,88] As shown below, our expert algorithm for substance identification (EASI) naturally accounts for correlations between peaks that derive from mass bias and spectral tilting.

MULTIVARIATE METHODS

By design, multivariate methods of spectral comparisons tend to account for the covariance or correlations between ion abundances in replicate spectra. Sigman and Clark examined the two-dimensional cross-correlations in replicate spectra of high explosives,[89] but the cross-correlations were examined as a function of a deliberate perturbation, i.e., differing collision energies, and the cross-correlations were not used to support compound identification, unlike the present work. In mass spectrometry applications, principal component analysis (PCA) has been used in two main ways: (1) to resolve or deconvolute mass spectra of mixtures,[66,68,70,90] and (2) to relate a spectrum to other classes or structures within a database.[74,91−96] The latter approach has also been used in conjunction with discriminant analysis and binary classification algorithms to enable the classification of spectra to known identities.[97−107] Finally, machine learning and artificial intelligence methods have existed since the early 1970s,[108] and they continue to be explored as methods to both identify known compounds in a library and to propose structures for compounds that are not in a library.[47,57,78,79,109−115] Whereas the predictive power of sophisticated computational techniques is likely to continually advance, very few of the articles described so far tackle the difficult problem of discriminating between spectrally similar compounds collected on different instruments and without reference spectra from those instruments.

The ability to resolve spectrally and structurally similar compounds, like isomers, using their EI-MS spectra has traditionally relied on the knowledge that certain ions are spectrally more unique or more valuable for discriminating isomers than others. For example, strongly correlating ions are thought to provide near-constant ratios of abundances that can be used to help discriminate between structurally similar compounds like positional isomers[94,99] and cocaine diastereomers.[116−118] Indeed, in the 1960s Tal'roze and Raznikov showed that if the correct ion ratios are taken into account, just three or four pairs of ratios are enough to resolve hundreds of compounds successfully.[119,120] In the case of cocaine, although the GC retention times can usually resolve the four major diastereomers of allococaine, cocaine, pseudococaine, and pseudoallococaine, mass spectral differentiation of the diastereomers within an instrument has been proposed through the ratios of peak abundances at m/z 94:96 and m/z 152:150.[116−118] However, the tactic of finding and examining specific pairs of ions to differentiate isomers of drugs[94,99] is a time-consuming approach that is not readily scalable to all isomers of all drugs or compounds. Analysts need an objective algorithm that is sufficiently flexible in identifying the ion ratios or correlations that are most characteristic or most discriminating for each substance without human intervention.

EXPERT ALGORITHM FOR SUBSTANCE IDENTIFICATION

We propose a new paradigm for mass spectral identification that does not require spectral similarity between the questioned spectrum and any of the reference spectra of the same substance to make a reliable identification. The described approach follows several guiding philosophies of valid computational methods,[78] including that (1) it be rooted in a solid scientific/mathematical basis, (2) it should have as few user-definable parameters as possible, (3) if possible, parameters should be learned from the data, and (4) if user-definable parameters are unavoidable, there should be very clear instructions on how to set these parameters depending on the experimental setup. As we will show, EASI can correctly discriminate known positives and known negatives even when known negatives are more similar to an exemplar or consensus spectrum[121] than other known positives, and EASI outperforms other algorithms even when data are collected on different types of mass analyzers. In this first demonstration, we show that EASI can reliably resolve a drug like cocaine from its diastereomers, even in the absence of chromatographic information.

The foundational basis and proof-of-concept of EASI is described in two parts. Part 1, here, describes how the kinetics of unimolecular fragmentation can be empirically modeled in a retrospective and passive manner from any existing database of replicate spectra of a reference material. Such replicates already exist in most crime laboratory settings because most jurisdictions already require spectra of standards to be collected contemporaneously with casework samples. Part 1 also describes how to assess and use the correlated variance in the replicates to build empirical models with which to compare the measured values. In part 2, the spectral similarity between modeled ion abundances and measured ion abundances is assessed in a variety of ways to enable binary classification using receiver operating characteristic (ROC) curves[13,34,122,123] to discriminate between cocaine and its diastereomers, selectively and confidently, and even when the spectra were collected decades apart and on a variety of different types of mass spectrometers.

https://doi.org/10.1021/jasms.3c00089
J. Am. Soc. Mass Spectrom. 2023, 34, 1248−1262

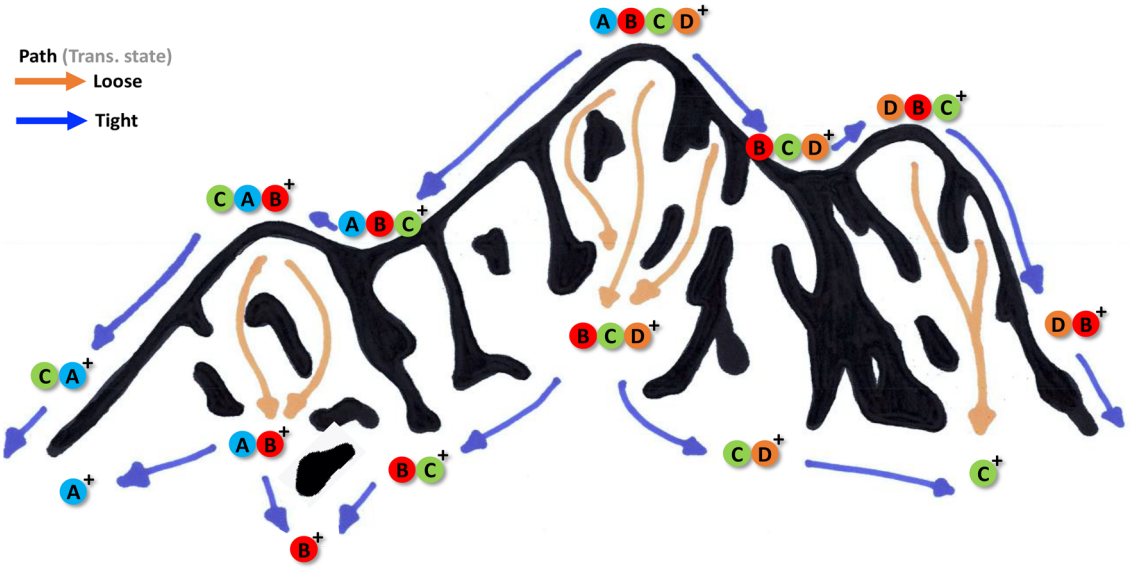


Figure 1. Schematic visualization of different potential energy pathways down an energy landscape. Tight transitions states can be thought of as narrow paths that take time to traverse and tend to involve rearrangements. Loose transition states can be thought of as wide-open pathways falling sharply off the mountain; they are faster and more numerous but can only be accessed from higher states.

Because almost all the questioned spectra of cocaine can be resolved from those of its diastereomers, which are the most likely to cause false positives, we assume the questioned cocaine spectra can easily be discriminated from all other known negatives. Such verification and validation is easily supported by conventional algorithms.

MATERIALS AND METHODS

A drug standard mixture analyzed by Laboratory 1 contained methamphetamine (1500 ppm), cocaine (1500 ppm), and hydromorphone (2000 ppm). Methamphetamine and hydromorphone were supplied by Sigma-Aldrich (St. Louis, MO, USA), cocaine was supplied by Mallinckrodt (St. Louis, MO, USA), and the methanol solvent was supplied by Alfa Aesar (Haverhill, MA, USA). The drug standard mixture analyzed by Laboratory 2 was comprised of ecgonine methyl ester (5050 ppm), cocaine (5050 ppm), 6-monoacetylmorphine (6-MAM) (5700 ppm), diacetylmorphine (DAM) (5700 ppm), and fentanyl (5200 ppm). The ecgonine methyl ester and cocaine were supplied by Sigma-Aldrich, the 6-MAM, DAM, and fentanyl were supplied by Lipomed (Cambridge, MA, USA), and the methanol solvent was supplied by Fisher Scientific (Waltham, MA, USA). The training set and test set are composed of data from abundant chromatographic peaks, so they represent somewhat ideal conditions that will not always be realized in casework. Future work, currently in progress, will consider the influence of absolute signal intensity on the performance of EASI.

An Agilent Technologies 7890 GC-5977 MS with a 12 m × 200 μm × 0.33 μm HP-5 (5% phenyl-methylpolysiloxane) column (Agilent J&W Columns, Santa Clara, CA, USA) was used by Laboratory 1. The GC method involved a 1 μL injection volume into a 220 °C injection port and a 100:1 split ratio. The initial oven temperature was held at 80 °C for 1.5 min, before being ramped to 270 °C at 50 °C/min and then held for 1.67 min. The method also included a second ramp to 290 °C with a 35 °C/min ramp rate and a 2.7 min hold. Helium was used as the carrier gas with a 1 mL/min flow rate. The transfer line temperature was set to 290 °C. The mass spectrometer scan range was m/z 30−650, with a 0.80 min solvent delay and a scan rate of 2852 Da/sec. The EI source temperature was 230 °C and the quadrupole temperature was 150 °C.

Laboratory 2 also used an Agilent Technologies 7890 GC-5977 MS; however, Lab 2 used a 30 m × 250 μm × 0.25 μm DB-5MS (5% phenyl-methylpolysiloxane) column (Agilent J&W Columns). The GC-MS method included a 0.2 μL injection volume into a 280 °C injection port and a 20:1 split ratio. The initial oven temperature was 80 °C, which was ramped to 300 °C with 30 °C/min ramp rate and then held for 9 min. The helium carrier gas was set to a flow rate of 0.684 mL/min. The transfer line temperature was set to 280 °C. The mass spectrometer scan range was m/z 40−500, with a 2 min solvent delay and a scan rate of 1472 Da/s. The EI source temperature was 230 °C and the quadrupole temperature was 150 °C. Most of the replicate spectra of cocaine and its diastereomers from the NIST archive come from a select number of laboratories dating back to the 1980s, including the NYC Police Laboratory (P. Shah), the NIST MS Data Center, the Defense and Civil Institute for Environmental Medicine, Canada (J. Zamecnik), the Georgia Bureau of Investigation (P. Price), and the Virginia Department of Forensic Science. Additional laboratory sources for cocaine spectra in the SWGDRUG database are provided on the SWGDRUG website.[124] Specific instruments and operating conditions of the database spectra are not known to us.

RESULTS AND DISCUSSION

Kinetic Basis for General Linear Modeling. When a particular organic molecule is ionized using electron- or photoionization, the observed branching ratios, that is, the abundance of peaks in the resulting mass spectrum, is determined by four main factors:[125−127] (1) the internal energy distribution of the molecule prior to ionization, (2) the excitation energy distribution accompanying ionization, (3) the apparent reaction time or observation time that is specific to the apparatus, and (4) mass bias and spectral distortion caused by ion optics and instrument conditions.[88,128] Branching ratios may be further affected by collisions between activated ions and residual gases en route to detection.

https://doi.org/10.1021/jasms.3c00089
J. Am. Soc. Mass Spectrom. 2023, 34, 1248−1262

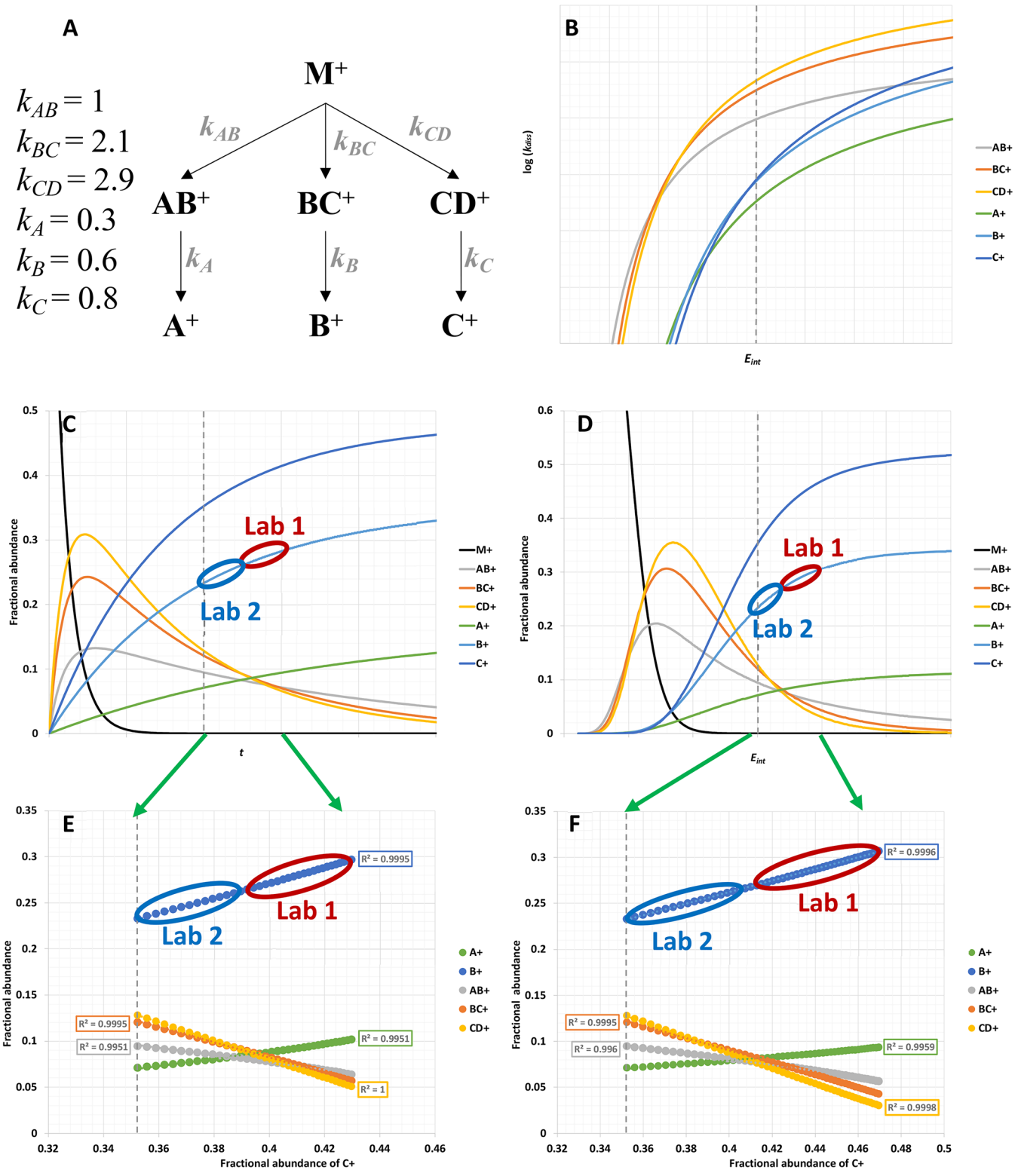


Figure 2. Theoretical modeling to show how changes in fractional ion abundances as a function of internal energy and observation time can be linearly extrapolated between instruments: (A) modeled system and relative rates of fragmentation at the vertical dashed line in each panel; (B) modeled log(k_{diss}) versus internal energy at a fixed reaction time; (C) breakdown curves as a function of time at a fixed internal energy; (D) breakdown curves as a function of internal energy at fixed time; (E) ion abundances versus the abundance of C^+ over the range of times in panel C; and (F) ion abundances versus the abundance of C^+ over the range of internal energies in panel D. One grid width in Figures B, C, and D spans a 33% increase in internal energy and 50% increase in reaction time relative to the dashed line for the conditions in panel A.

The relationship between the internal energy distribution of an activated ion and the branching ratios of its product ions can be explained by classical statistical models such as the quasi-equilibrium theory (QET)[127] or the Rice−Ramsperger−Kassel−Marcus (RRKM) theory of unimolecular dissociation.[127,129−133] These theories pose that the reaction rate of a

https://doi.org/10.1021/jasms.3c00089
J. Am. Soc. Mass Spectrom. 2023, 34, 1248−1262

particular pathway is determined by the entropy and enthalpy of activation of the transition state,[129,134,135] and they are sufficiently reliable to enable EI fragmentation patterns to be accurately modeled and reasonably well predicted from first principles.[127,136] Bauer and Grimme provide an excellent account of the rich history in this area.[136]

One way to appreciate the mathematical relationship between the dissociation rate and the density of transition states is through the Kassel equation,[137,138] which shares general trends with the more-sophisticated QET and RRKM theories but is easier to understand.[125,135] The flow of different ions through multicoordinate transition space can be thought of as a precursor ion having many different pathways down a molecular landscape, as shown in Figure 1. In this analogy, certain elevations, and therefore certain pathways, are not accessible if the excitation energy is too low that the precursor can only ascend part-way up the mountain.

The Kassel equation (eq 1) assumes that the rate of a reaction as a function of energy $k(E)$ is the product of the frequency factor of a certain energy level and the probability of populating the energy level:

$$k(E) = \nu\left(\frac{E - E_0}{E}\right)^{3N-7} \quad (1)$$

where ν is an entropic factor or frequency factor, proportional to $e^{-E_0/kT}$, that describes the tightness of the transition state. The term $\left(\frac{E-E_0}{E}\right)$ describes the fractional probability of populating an energy level E above the critical energy E_0. As the internal energy of the precursor ion increases, the fraction $\left(\frac{E-E_0}{E}\right)$ approaches unity. The term N is the number of atoms in the precursor, so the power term $3N-7$ is the number of vibrational degrees of freedom in a nonlinear precursor (not including the reaction co-ordinate).

A schematic visualization of Kassel's equation is shown in a potential energy hypersurface in Supplemental Figure S2. Kassel's equation explains how the dissociation rate of a reaction increases when the entropy is more favorable, when the activation energy increases, or when the number of atoms is smaller. The corollary is that larger ions require larger internal energies to observe even the lowest energy rearrangements, a phenomenon which has hindered the effectiveness of CID of high-mass ions in top-down proteomics.[139−141]

When millions of activated precursor ions are formed in an EI source at any given instant, they start with a distribution of internal energies, or elevations up the mountain. Ions can either follow loose transitions, represented by steep, fast, and unspecific pathways down the mountain, or tight transitions, represented by less steep, slower, and specific pathways down the mountain. The ions end their paths with frequencies of occurrence in proportion to their statistical probabilities. At short observation times, the only fragments that can be observed are those that transition through fast or loose transition states. At longer observation times, fragments occurring through longer or tighter transition states become more prominent. Below, we illustrate how these basic principles of QET/RRKM transition state theory lead to approximately linear fragment-fragment correlations that can be effectively fitted with general linear models to extrapolate the kinetic-based behavior between instruments. To be clear, QET/RRKM theory does not result in a mathematically rigorous proof of linear branching ratios. Instead, basic modeling is used to show that when the activation energy vastly exceeds the appearance energies of the different pathways, then modest changes of ∼30% in the internal energy or 50% in the reaction time will provide some approximately linear relationships between at least several pairs of ions. In practice, the method of mixed stepwise selection of general linear modeling can effectively identify and employ the ions that best correlate with one another among the replicate spectra to build the general linear models (GLMs). Neither human selection nor knowledge about the structural relationship between different ions is required for successful implementation of EASI.

Simulated Kinetics and Branching Patterns. Consider a precursor molecular ion M^+ with the generic structure $ABCD^+$. QET/RRKM theory can apply to both odd- and even-electron ions, so a radical is not shown for simplicity. Figure 2 shows a hypothetical branching pattern of a simple molecule that includes both competitive and sequential fragmentations. For simplicity, the modeled rates have been normalized to the rate of k_{AB} at a given arbitrary internal energy (E_{int}) and observation time (t), as provided in Figure 2A. The modeling is based on a hypothetical precursor with 70 atoms, arbitrary activation energies between 1.6 and 3 eV and arbitrary frequency factors. The modeling further assumes that the internal energy of a precursor is distributed between the free energy of the pathway and the internal energy of the fragment ions. The fractional ion abundances (assuming $[M^+]_{t=0} = 1$) as a function of time (t) were derived from a combination of branching rates and consecutive rates according to the following examples:[125,134−136]

$$[M^+]_t = e^{-k_{tot}t} \quad (2)$$

where $k_{tot} = k_{AB} + k_{BC} + k_{CD}$.

$$[AB^+]_t = \frac{k_{AB}}{k_{tot}}(e^{-k_A t} - e^{-k_{AB}t}) \quad (3)$$

$$[A^+]_t = \frac{k_{AB}}{k_{tot}}e^{-k_{tot}t} - \frac{k_{AB}}{k_{tot}}e^{-k_A t} - e^{-k_{tot}t} \quad (4)$$

Figure 2A shows the modeled fragmentation pathway and Figure 2B shows the rate of change of $\log(k_{diss})$ versus internal energy E_{int} for the different pathways. The fragments in the key are listed in ascending order of activation energy and descending order of tightness of the transition state. Therefore, AB^+ is the lowest energy rearrangement product and is observed at the lowest appearance energies.[134,135]

In Figure 2C−F, the fragment C^+ has the highest appearance potential but the loosest transition state, so its rate increases more quickly with internal energy than the other pathways. The regions labeled "Lab 1" indicate a hypothetical range of measurements observed by one laboratory based on the variance in their data, which is characterized by the specific geometry of the lab's instrument, the type of mass spectrometer, and their specific operating conditions. If Lab 2 uses a different geometry of instrument, such as from a different vendor, or uses different operating conditions, then the relative ion abundances observed in Lab 2 will differ from Lab 1 because of the different internal energies and observational timeframes of fragmentation. For example, in Figure 2C, the mean fractional abundance in Lab 1 for the fragment B^+ is 0.270 ± 0.015% (95% confidence interval) and in Lab 2, the mean fractional abundance is 0.245 ± 0.015% (95% confidence interval). A Student's t-test could easily show

https://doi.org/10.1021/jasms.3c00089
J. Am. Soc. Mass Spectrom. 2023, 34, 1248−1262

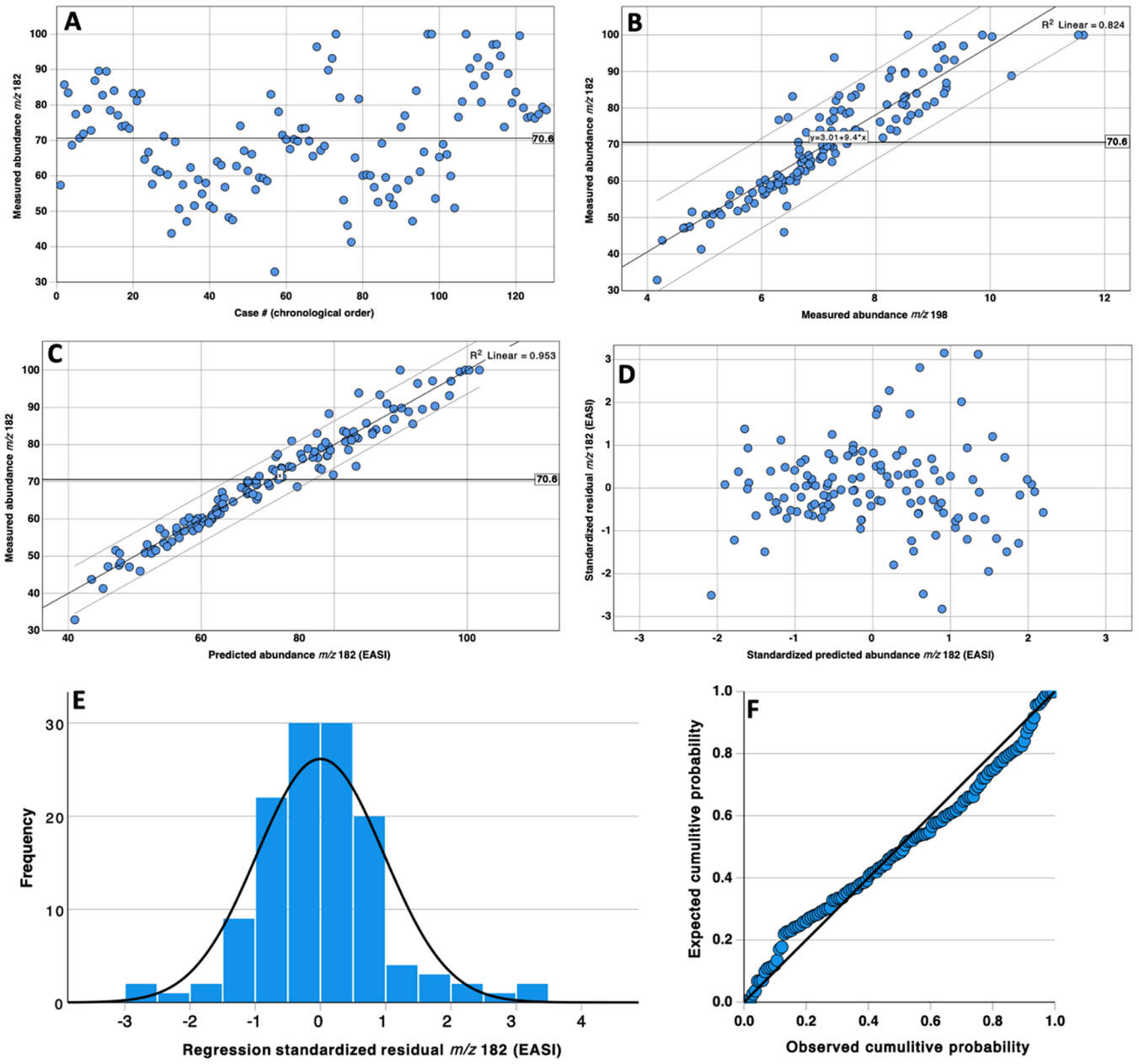


Figure 3. Scatter plots of measured and predicted ion abundances for m/z 182 for cocaine: (A) normalized abundance of m/z 182 relative to the base peak in 128 spectra (cases) collected over 6 months; (B) scatter plot of the normalized abundance of m/z 182 versus the normalized abundance of m/z 198 in the same data; (C) scatter plot of the normalized abundance of m/z 182 versus the EASI-predicted abundances using the coefficients shown in Table 2; (D) scatter plot of the standardized residuals versus the standardized predicted abundances based on the 128 predictions in panel C; (E) frequency distribution plot of the standardized residuals of the 128 predictions in panels C and D; and (F) P−P plot of the standardized residuals in panels C and D.

that the abundance of B^+ is significantly different at $p < 0.05$, and any "good" algorithm that requires spectral similarity as a modus operandi should find most replicates from Lab 2 are significantly different from most replicates of Lab 1. Other fragment ions may also be significantly different between the two laboratories. These differences between instruments are limiting factors in the effectiveness of algorithms that use spectral similarity as a mechanism for compound identification.

In contrast to spectral similarity methods that are typically used to identify substances,[13,37,42,71] the linear regression lines in Figure 2E,F indicate that the linear regression equations for Lab 1's data can be extrapolated to predict ion abundances in regions of experimental space that are beyond the natural variance captured in Lab 1's data. In other words, Lab 1 could extrapolate the trend in its data to predict ion abundances for data collected in Lab 2. Such capabilities are generally not possible when the variance within a training set does not capture the variance of the test or validation set, as demonstrated by the PCA plot in Supplemental Figure S3.

For an analyst in Lab 1 to make accurate predictions about fragment ion abundances in a spectrum of M^+ in Lab 2, Lab 1 would only need to know the values of the covariate ion abundances (i.e., a measured spectrum from Lab 2) and the coefficients derived from the various regression models. Lab 1 would not need to know the source of the variance nor conduct any corrections or adjustments to Lab 2's spectra to minimize

https://doi.org/10.1021/jasms.3c00089
J. Am. Soc. Mass Spectrom. 2023, 34, 1248−1262

Table 1. Summary of Bivariate Pearson Correlations for the 20 Most Abundant Fragments of Cocaine[a]

	Covariate m/z																			
m/z	**42**	**51**	**55**	**67**	**68**	**77**	**81**	**82**	**83**	**94**	**96**	**97**	**105**	**122**	**152**	**182**	**183**	**198**	**272**	**303**
42	1.000	.937	.953	.615	.574	.519	.040	.146	.165	.518	.599	.712	.477	.427	.300	.433	.413	.278	.346	.377
51	.937	1.000	.892	.706	.635	.657	.005	.138	.074	.595	.597	.658	.591	.480	.317	.432	.392	.314	.369	.397
55	.953	.892	1.000	.593	.474	.421	.021	.171	.101	.417	.510	.627	.371	.371	.266	.369	.346	.216	.288	.318
67	.615	.706	.593	1.000	.772	.823	.286	.270	.051	.756	.704	.569	.770	.502	.321	.405	.355	.259	.300	.344
68	.574	.635	.474	.772	1.000	.897	.216	.271	.143	.873	.754	.687	.863	.562	.361	.486	.464	.306	.317	.343
77	.519	.657	.421	.823	.897	1.000	.183	.205	.047	.922	.758	.646	.954	.562	.390	.517	.457	.353	.395	.429
81	.040	.005	.021	.286	.216	.183	1.000	.489	.484	.213	.247	.114	.152	.136	.042	-.079	-.078	-.169	-.204	-.225
82	.146	.138	.171	.270	.271	.205	.489	1.000	.542	.136	.169	.141	.146	-.070	-.123	-.263	-.284	-.398	-.455	-.437
83	.165	.074	.101	.051	.143	.047	.484	.542	1.000	.090	.296	.320	.048	-.032	.048	.041	.051	-.072	-.140	-.159
94	.518	.595	.417	.756	.873	.922	.213	.136	.090	1.000	.864	.757	.959	.719	.551	.645	.611	.488	.505	.532
96	.599	.597	.510	.704	.754	.758	.247	.169	.296	.864	1.000	.881	.816	.648	.606	.739	.730	.582	.576	.581
97	.712	.658	.627	.569	.687	.646	.114	.141	.320	.757	.881	1.000	.702	.593	.588	.761	.756	.608	.606	.618
105	.477	.591	.371	.770	.863	.954	.152	.146	.048	.959	.816	.702	1.000	.682	.507	.614	.558	.471	.470	.508
122	.427	.480	.371	.502	.562	.562	.136	-.070	-.032	.719	.648	.593	.682	1.000	.690	.699	.666	.640	.523	.533
152	.300	.317	.266	.321	.361	.390	.042	-.123	.048	.551	.606	.588	.507	.690	1.000	.792	.779	.760	.640	.606
182	.433	.432	.369	.405	.486	.517	-.079	-.263	.041	.645	.739	.761	.614	.699	.792	1.000	.962	.908	.868	.839
183	.413	.392	.346	.355	.464	.457	-.078	-.284	.051	.611	.730	.756	.558	.666	.779	.962	1.000	.892	.865	.834
198	.278	.314	.216	.259	.306	.353	-.169	-.398	-.072	.488	.582	.608	.471	.640	.760	.908	.892	1.000	.827	.762
272	.346	.369	.288	.300	.317	.395	-.204	-.455	-.140	.505	.576	.606	.470	.523	.640	.868	.865	.827	1.000	.963
303	.377	.397	.318	.344	.343	.429	-.225	-.437	-.159	.532	.581	.618	.508	.533	.606	.839	.834	.762	.963	1.000

[a]Abundances were normalized to the base peak before analysis. $N = 128$ spectra over 6 months from an operational crime laboratory (Lab 1). The two largest bivariate Pearson correlations in each row are shaded gold.

the residual errors between the predicted and measured ion abundances. One caveat is that Lab 1's spectral variance may be dominated by changes in excitation energy whereas a Lab 2's spectra might differ from Lab 1 primarily because of a difference in apparent observation times. In such cases, if the rate of change in ion abundances as a function of time and energy are disparate, then the linear models based on changes in energy may not make accurate predictions about ion abundances affected by changes in time. Such issues could be answered and solved by using a training set that incorporates data from all the instruments of intended deployment.

Figure 2E,F indicates that, whether the variance between Labs 1 and 2 are caused by differences in internal energies or observation times, there are likely to be some strong linear relationships that relate the two domains of variance. EASI outlines a basic framework to help identify the ions that best explain one another's abundance and thereby enable robust and accurate predictions between laboratories. Here, we use stepwise general linear modeling, but many other solutions to interlaboratory comparisons are conceivable based on the same underlying assumption of linear correlations between ion abundances of replicate spectra.

To gain an appreciation of the natural variance of correlations among measured product ions in replicate spectra, Figure 3A shows the relative abundance of m/z 182 versus case number for 128 replicate analyses of a cocaine standard in an operational forensic laboratory over a 6-month period.[28] The database includes an average of about one cocaine spectrum per business day. Our previous analysis of the same data included a brief observation of the correlations between ion abundances in these replicate spectra, but the previous study did not address the source of the variance nor how to take advantage of the correlations to enable spectral identifications.[28] When the normalized abundance of the peak at m/z 182 of cocaine is plotted as a function of the case number, the abundances vary over the wide range of ~34−100% (Figure 3A). In cases where m/z 182 was the base peak, the abundance of m/z 82 was as low as 78%. Again, most past and present algorithms assume this scatter is random[48,142] and/or that this variance can only be controlled using "properly tuned" instruments.[38,55] However, the bivariate plot in Figure 3B shows that the variance in the normalized abundance of m/z 182 is not random and that, for example, ~ 82% of the variance in its normalized abundance can be explained by the normalized abundance at m/z 198. This is to say that the abundance at m/z 182 correlates far better with the abundance at m/z 198 than it does with the peak at m/z 82, which is typically the base peak.

Table 1 shows the bivariate correlations for the 20 most abundant fragments in the training set of 128 cocaine spectra collected over 6 months in Lab 1. Admittedly, the choice to model the 20 most abundant ions is somewhat arbitrary. The number of variables could be optimized and validated in future work and with different substances. In this case, the choice to include 20 ion abundances deliberately enabled the inclusion of the fragments at m/z 94 and m/z 152, which are known to be important for the discrimination of cocaine from its diastereomers, like pseudococaine.[117,118] The introduction includes many examples of algorithms that report only a slight loss in performance when using smaller fractions of the measured peaks in spectral comparison algorithms.[11]

The statistical significance of each cross-correlation is provided in Table S1. Table 1 shows that, in general, ions that appear close in m/z value to one another tend to correlate more strongly than with ions of disparate m/z values. This observation can be thought of as a mass bias or tuning effect, which causes ions of similar m/z to correlate. Exceptions are m/z 105 and m/z 77, which share a strong correlation by virtue of the close mechanistic and kinetic relationships between them, i.e., m/z 77 is a direct cleavage product of −CO (28 Da) from m/z 105.[116−118] As shown in Table 1, the normalized abundance of m/z 198 visualized in Figure 3B is one of many ions that correlate strongly with the normalized abundance at m/z 182, so several other ions could also serve as independent variables with

https://doi.org/10.1021/jasms.3c00089
J. Am. Soc. Mass Spectrom. 2023, 34, 1248−1262

Table 2. Summary of the Unstandardized Coefficients for 20 General Linear Regression Models Using the Abundance of Each *m/z* Value as a Dependent Variable and the Remaining 19 Abundances as Possible Covariates*

Dependent *m/z* value	Unstandardized coefficients for covariate *m/z* values																				
	β_0	42	51	55	67	68	77	81	82	83	94	96	97	105	122	152	182	183	198	272	303
42	6.43		0.96	**1.78**		0.97	-0.11		-0.26	0.20			0.88						-0.52		
51	0.47	**0.36**		0.49		-0.41	0.19	-0.50		0.16	-0.17		-0.27		0.48		-0.10		0.67	0.24	
55	-1.98	**0.22**			0.30				0.09	-0.12				-0.06			0.012				
67	-1.02	-0.05		0.35			**0.05**	0.25													
68	0.89						**0.05**				0.04		0.12								-0.03
77	-12.4	-0.57		1.47		2.88		1.52		-0.31	0.35			**0.87**	-2.09						
81	0.72		-0.06		**0.31**				0.04	0.11					0.17		0.01				
82	**76.2**				0.76					0.46			0.46			0.88			-0.62	-0.64	
83	-1.38				-0.77			1.25	0.27				**0.59**		-0.32						
94	4.56					0.99						0.42		0.42	0.52						
96	0.23		-0.07		0.73					0.20	0.15		**0.60**					0.31			
97	-7.35	0.09			-0.46				0.10			**0.29**		0.02				0.31			
105	-0.25			0.48			**0.42**	-0.93			0.42	0.32			0.89						
122	1.36	-0.06	0.23				**-0.13**	0.40		-0.10		0.16	-0.19	0.10			0.06			-0.21	
152	-1.96					-0.12			0.03						0.16		**0.03**				
182	-16.6[a]						0.13[a]			0.38[a]						1.88[a]		**4.44**[a]	2.72[a]		0.34[a]
183	-2.06					0.34	-0.04					0.09					**0.08**		0.14	0.15	
198	9.57					-0.15			-0.08							0.36	**0.08**				
272	0.38														-0.12				0.34		**0.27**
303	3.39						-0.08	-0.55						0.18					-0.65	**3.11**	

*Bold underlined font indicates the most significant standardized coefficient in each regression model. $N = 128$ spectra were used as the training set with the 20 most abundant peaks selected. [a]As an example of how to use these coefficients, these coefficients are employed in eq 6 to predict the abundance of m/z 182 in any given query spectrum.

which to model or predict the abundance at m/z 182 as the dependent variable.

Extending this idea further, the abundance of m/z 182 as a dependent variable can be modeled or predicted using a general linear model by employing multiple covariates in a single model, as shown in eq 5:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \ldots + \beta_n x_n + \varepsilon \quad (5)$$

where y is the abundance of the dependent variable (e.g., m/z 182), β_0 is the y-intercept, β_1 through β_n are the covariate coefficients, x_1 through x_n are the covariate ion abundances (e.g., m/z 198 and m/z 183), and ε is the residual error. Conceptually, general linear modeling (GLM) can be thought of as a weighted average of more than one linear regression model ($y = \beta_0 + \beta_1 x_1 + \varepsilon$). GLM is a classical statistical technique that was elegantly demonstrated by Sir Francis Galton in 1886 when he showed that the height of an adult, the dependent variable, could be accurately modeled or predicted using the heights of his/her biological parents, the covariates.[143] In an application vaguely similar to the one described here, GLM has also been used to estimate missing values (not m/z abundances) from a failed sensor to help stabilize the behavior of different multivariate models in a process-control setting.[144] General linear modeling has also been used to predict GC retention indices as the dependent variable from a variety of molecular properties as covariates.[145]

In Figure 3C, the abundance of m/z 182 as the dependent variable was modeled in IBM's SPSS software (version 28.0) using stepwise development of a general linear model. Covariates were added when their contribution to the model was significant at $F \leq 0.05$ and removed when their contribution was insignificant at $F \geq 0.1$. Of the 20 developed models, the models varied from as few as three covariates (not including the y-intercept) for m/z 272 to as many as 12 covariates to effectively model m/z 51. The unstandardized coefficients in the 20 GLM models are provided in Table 2. The fact that the stepwise models never require all 20 ions to maximize the extent of explainable variance is an indication that satisfactory performance of GLM could probably be obtained using fewer than 20 ion abundances at the onset. Again, the focus of the current manuscript is a proof-of-concept rather than complete optimization.

As an example of how to use this table, the abundance of m/z 182 ($\hat{A}_{182}$) in any cocaine spectrum can be predicted using the remaining ion abundances in the spectrum using the equation:

$$\hat{A}_{182} = -16.6 + 0.13A_{77} + 0.38A_{83} + 1.88A_{152} + 4.44A_{183} + 2.72A_{198} + 0.34A_{303} \quad (6)$$

The results of 128 such predictions using this model are shown against the measured values for the training set in Figure 3C. The coefficient of determination (R^2) of the regression line shows that this general linear model explains more than 95% of the variance in the measured abundances of m/z 182. Therefore, more than 95% of the variance in the abundance of m/z 182 in the 128 spectra is not random and is explainable through covariance mapping. Figure 3D–F shows different ways of assessing the distribution of the residual differences between the modeled (predicted) abundances and the measured abundances at m/z 182 for the training set. In short, the residuals follow a normal distribution, are not significantly biased or skewed, and do not display problematic kurtosis. The residuals were also assessed for the remaining 19 models, and only 4 of the 20 models contained statistically significant skew. Seventeen of the 20 models contained kurtosis greater than the margin of error (0.425), and most had kurtosis measures >3, which indicates that the residuals generally displayed leptokurtosis with wider tails than Gaussian. Therefore, several of the observed P–P plots

https://doi.org/10.1021/jasms.3c00089
J. Am. Soc. Mass Spectrom. 2023, 34, 1248−1262

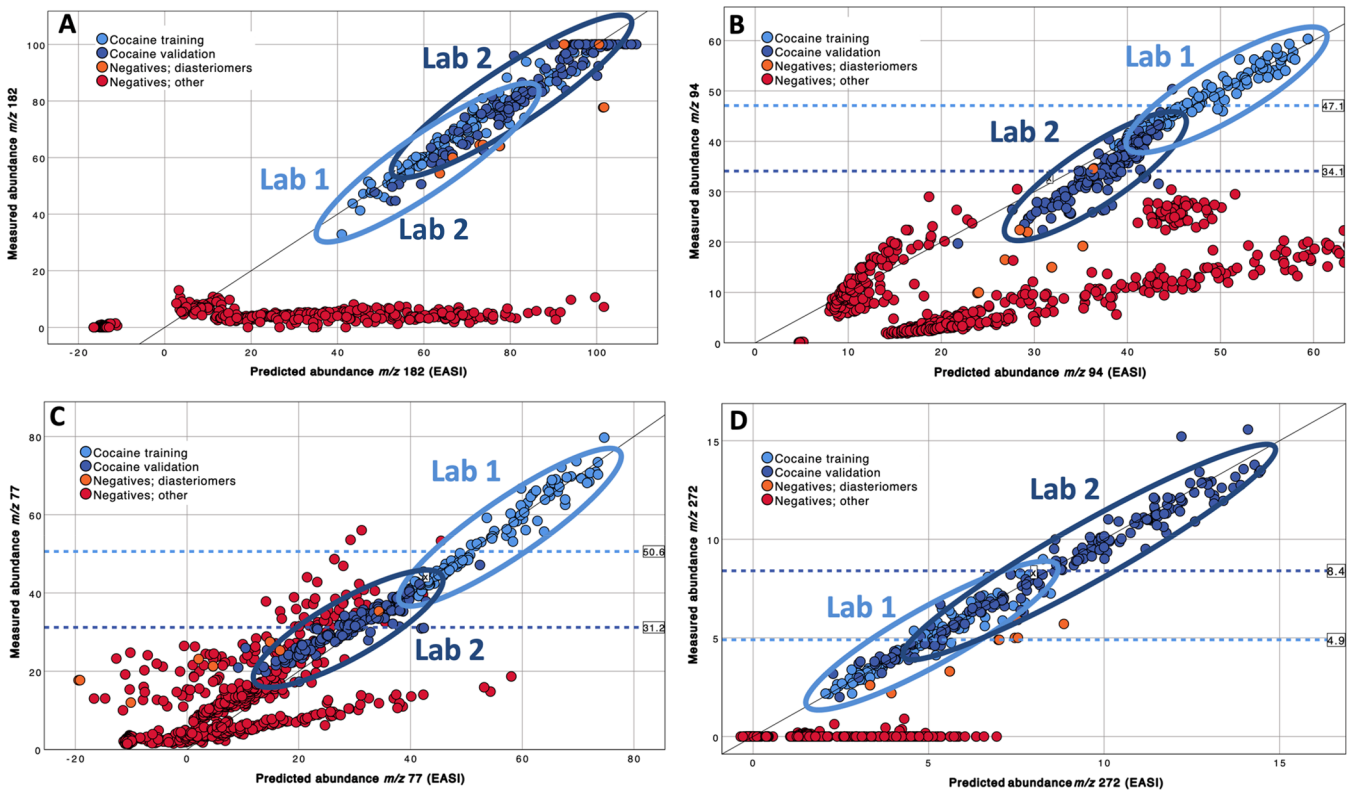


Figure 4. Scatter plots of measured and modeled/predicted values for a few selected ions of cocaine: (A) m/z 182, (B) m/z 94, (C) m/z 77, and (D) m/z 272. Horizontal lines show the mean abundances for the training set from Lab 1 (light blue) and validation set from other laboratories (dark blue). The line $y = x$ refers to the ideal case of no residual error in predictions.

showed some deviation from perfect normality, but the deviations were mild and considered satisfactory as a proof-of-concept of this approach.

Future work should continue to address the validity of the assumptions necessary to support the proposed statistical modeling. For example, given that normalized ion abundances can never be negative, the measured distributions might act more Poisson than Gaussian, so modeling methods could be adjusted accordingly. As a final measure of validity, we also performed a correlation analysis on the residuals (ε) that resulted from GLM. Covariance mapping of the residuals of the training set predictions showed that only ~10% of the 190 pairwise comparison tests between the different m/z channels showed any significant correlation. Therefore, GLM effectively explained and removed almost all the cross-correlations that existed in the normalized abundances of the replicate spectra in the training set.

So far, the discussion has shown that the branching ratios of the 128 cocaine spectra in the training set contain ion abundances that are strongly linearly correlated, as approximated over modest changes in energy and time by QET/RRKM theories of unimolecular fragmentation. Also, that GLM is a reasonably valid approach that can explain more than 90% of the variance in normalized ion abundances. The next step is to demonstrate that linear models built on the training set from one instrument can be extrapolated to spectra collected on different instruments.

To demonstrate such extrapolation, the general linear models built on the training set of 128 cocaine spectra were applied to several groups of spectra: (1) a validation set that consisted of 120 validation cocaine spectra from a second crime laboratory (Lab 2); (2) 55 validation cocaine spectra from the NIST archive; (3) 706 replicate known negative spectra of ecgonine methyl ester, fentanyl, heroin, hydromorphone, and methamphetamine; and (4) a total of 10 replicate known negative spectra of the four diastereomers of cocaine from the NIST archive, including allococaine, pseudococaine, and pseudoallococaine. Bivariate plots of measured abundances versus predicted abundances for a select number of fragments of cocaine are provided in Figure 4 and Supplemental Figures S4–S8. Summary statistics for the four groups of spectra are also provided in Table S2.

Figure 4A is a scatterplot of the measured abundance of m/z 182 versus the predicted abundance of m/z 182 for all 1019 spectra in the database. The general linear model is the one shown in eq 6, but in this case the model is applied to all spectra in the database, including all known positives (KPs) and known negatives (KNs). The box and whisker plots in Supplemental Figure S9A show the same measured values in Figure 4A in a manner that makes the distributions of measured values easier to compare. Note that the KN diastereomers in orange provide measured abundances of m/z 182 that are closer to the mean of the training set (i.e., the consensus spectrum[57,121]) than most of the 175 KP cocaine validation spectra. For example, the mean abundances for m/z 182 are 70.6% for the KP training set and 72.7% for the KN diastereomers, but 88.4% for the KP validation set.

Most of the other KNs provide abundances for m/z 182 that are <10% relative abundance, so they are easily dismissed as not behaving like cocaine. Therefore, whereas all the measured abundances of m/z 182 in the spectra of cocaine and its diastereomers are notably different from the known negatives in

https://doi.org/10.1021/jasms.3c00089
J. Am. Soc. Mass Spectrom. 2023, 34, 1248−1262

the database, the relative abundances of m/z 182 of cocaine diastereomers are not significantly different (t-test, two-tailed, α = 0.05) from that of the cocaine training set.

Figure 4B is a scatterplot of the measured abundance of m/z 94 versus the predicted abundance of m/z 94 for all 1019 spectra in the database. Again, the general linear model was built on the training spectra of 128 cocaine spectra from Lab 1 and used the coefficients provided in the row for m/z 94 in Table 2. Box and whisker plots of the same data are shown in Supplemental Figure S9B. In this case, the training set has a mean abundance of 47.1% for m/z 94, the validation spectra of cocaine have a mean abundance of 34.1%, and the diastereomers have a mean abundance of 18.4%. The outlier in the diastereomers is a spectrum of pseudoallococaine, which has been shown to share spectral similarity with cocaine because of the steric similarity of the activated complexes.[117,118] Here, the measured ranges of the training and validation sets of cocaine are significantly different (one-way ANOVA, post hoc least significant difference (LSD): $p < 0.001$). However, the regression line in Figure 4B enables accurate predictions for abundances of m/z 94 in the test set, even though the measured values in the test set are outside most of those in the training set. For example, the smallest abundance for m/z 94 in the training set of Lab 1 was 35.1%, but more than half of the test set provided abundances below this threshold. The ability to extrapolate the general linear model in Figure 4B to fit the abundance measurements of known positives made on different instruments and in different laboratories is unique to this algorithm, and it demonstrates the reliance on statistical behavior predicted by approximations of QET/RRKM theory, as demonstrated in Figure 2E,F.

The scatter plots of measured versus modeled abundances for m/z 77 and m/z 272 in Figure 4C,D further support the contention that the modeled linear behavior in one laboratory can be used to make accurate predictions about ion abundances in other laboratories, even when the range of measured values is substantially different between the laboratories. Supplemental Figures S4−S8 provide additional examples of effective extrapolations for fragments at m/z 68, 105, 122, 152, and 303, respectively. Each model varies in its effectiveness at extrapolating between the two laboratories. In many of these figures, the predicted abundances for the diastereomers of cocaine have larger residual errors than most of the cocaine spectra.

Although the current manuscript only provides modeling data for cocaine, we have applied GLM to replicate spectra of various analytes on different types of instruments, including: (1) electron ionization spectra from GC-MS instruments, (2) electrospray ionization tandem mass spectrometry (ESI-MS/MS) data from a quadrupole time-of-flight instrument, and (3) direct analysis in real time (DART) MS/MS spectra from a triple-quadrupole mass spectrometer. In all cases, multivariate linear models provided similar figures of merit as presented here, such as explaining more than 90% of the variance in the normalized fragment ion abundances, and that the residual errors between modeled and measured abundances were distributed in an approximately Normal manner. Furthermore, the residuals using EASI were typically ∼4 times smaller than residuals between measured abundances and their corresponding centroid values of the training set. Therefore, the results for cocaine presented here can be considered a typical result rather than an outlier.

These results indicate that instead of focusing on one or two specific ion ratios to enable discrimination between a compound and a closely related structure (e.g., cocaine and allococaine),[116−118] GLM can help explain other sources of variance between cocaine and its diastereomers. There are many ways that one could use correlations between different pairs of ions in replicate spectra to establish whether the fragment ion abundances in a questioned spectrum follow the expected behavior of a substance or not. Several examples are described in part 2 of this manuscript.

CONCLUSIONS

This manuscript demonstrates that GLM is a reasonably robust and valid approach to model the branching ratios of replicate spectra of cocaine. The residual differences between measured and modeled abundances for the 20 most abundant fragments of cocaine are approximately normally distributed and uncorrelated with one another, unlike the input normalized abundances. Most importantly, the models can be extrapolated to make accurate and robust predictions of relative ion abundances for cocaine spectra collected on various types of mass analyzers in different laboratories dating back to the 1980s. One example for the peak at m/z 182 shows that the variance in normalized ion abundance ranges from 40 to 100% (70 ± 30% confidence interval) in the training set, but a simple general linear model with one constant and six terms enables the same ion abundances to be predicted with confidence intervals less than ±2%. The linear behaviors modeled in this work are predicted by QET/RRKM theories of unimolecular fragmentation that were developed in the 1950s, and they are modeled using an approach that has been in use since at least the 1880s. In the future, neural networks and machine learning algorithms could also be used to take advantage of the linear relationships predicted by QET/RRKM theory. Finally, the developed coefficients for cocaine provided in Table 2 can be considered reasonably valid for predicting the relative ion abundances within any 70 eV mass spectrum of cocaine on any mass spectrometer, in perpetuity. To apply EASI to future casework samples of cocaine, analysts would not need access to a database of hundreds of replicate spectra of cocaine to enable reliable spectral identification, they would only need to refer to a table containing fewer than 100 linear coefficients, as provided in Table 2. The companion manuscript describes how to use the GLM models as a binary classifier to enable effective discrimination between cocaine and any known negatives, including its diastereomers.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/jasms.3c00089.

3 tables and 9 figures, including: (1) normalized abundnaces for cocaine across a GC peak; (2) schematic visualization of a potential energy hypersurface; (3) summary statistics for 20 ion abundances of different groups of compounds; (4) PCA analysis results; (5) bivariate plots of GLM-predicted abundances versus measure abundances, and (6) box-and whisker plots of measured abundances for different data sets (PDF)

AUTHOR INFORMATION

Corresponding Author

Glen P. Jackson − *Department of Forensic and Investigative Science, West Virginia University, Morgantown, West Virginia*

https://doi.org/10.1021/jasms.3c00089
J. Am. Soc. Mass Spectrom. 2023, 34, 1248−1262

26506, United States; C. Eugene Bennett Department of Chemistry, West Virginia University, Morgantown, West Virginia 26506, United States; orcid.org/0000-0003-0803-6254; Phone: 304-293-9236; Email: glen.jackson@mail.wvu.edu

Authors

Samantha A. Mehnert − *Department of Forensic and Investigative Science, West Virginia University, Morgantown, West Virginia 26506, United States; C. Eugene Bennett Department of Chemistry, West Virginia University, Morgantown, West Virginia 26506, United States;* Present Address: Department of Chemistry, Purdue University, West Lafayette, Indiana 47907, United States

J. Tyler Davidson − *Department of Forensic and Investigative Science, West Virginia University, Morgantown, West Virginia 26506, United States;* Present Address: Department of Forensic Science, Sam Houston State University, Huntsville, Texas, 77340, United States; orcid.org/0000-0001-9932-8273

Brandon D. Lowe − *C. Eugene Bennett Department of Chemistry, West Virginia University, Morgantown, West Virginia 26506, United States*

Emily A. Ruiz − *C. Eugene Bennett Department of Chemistry, West Virginia University, Morgantown, West Virginia 26506, United States*

Jacob R. King − *C. Eugene Bennett Department of Chemistry, West Virginia University, Morgantown, West Virginia 26506, United States*

Complete contact information is available at: https://pubs.acs.org/10.1021/jasms.3c00089

Notes

The opinions, findings, and conclusions or recommendations expressed in this publication/program/exhibition are those of the authors and do not necessarily reflect the views of the Department of Justice.

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This project was supported by grant 15PNIJ-21-GG-04179-COAP, awarded by the National Institute of Justice, Office of Justice Programs, US Department of Justice. We also thank the National Science Foundation (NSF) for supporting undergraduate students in the REU program through CHE-1852369. We would like to thank Benny Lum of the Broward Sheriff's Office Crime Laboratory and Gina Nano of the University of Massachusetts Medical School of Medicine for providing the raw data. We also thank various colleagues for their feedback on the algorithm and FACSS/SciX for the 2021 Innovation Award for this work.

■ REFERENCES

(1) Bleakney, W.; Condon, E. U.; Smith, L. G. Ionization and Dissociation of Molecules by Electron Impact. *J. Phys. Chem.* **1937**, *41*, 197−208.

(2) Herzog, R. F. K.; Viehbock, F. P. Ion Source for Mass Spectrography. *Phys. Rev.* **1949**, *76*, 855−856.

(3) Beynon, J. H. Qualitative Analysis of Organic Compounds by Mass Spectrometry. *Nature* **1954**, *174*, 735−737.

(4) Beynon, J. H. The use of the mass spectrometer for the identification of organic compounds. *Microchim. Acta* **1956**, *44*, 437.

(5) Frerot, E.; Wunsche, L. 50 Years of Mass Spectrometry at Firmenich: A Continuing Love Story. *Chimia* **2014**, *68*, 160−163.

(6) Hoffmann, W. D.; Jackson, G. P. Forensic Mass Spectrometry. *Annu. Rev. Anal. Chem.* **2015**, *8*, 419−440.

(7) Kopka, J. Current Challenges and Developments in GC-MS based Metabolite Profiling Technology. *J. Biotechnol.* **2006**, *124*, 312−322.

(8) Kind, T.; Wohlgemuth, G.; Lee, D. Y.; Lu, Y.; Palazoglu, M.; Shahbaz, S.; Fiehn, O. FiehnLib: Mass Spectral and Retention Index Libraries for Metabolomics Based on Quadrupole and Time-of-Flight Gas Chromatography/Mass Spectrometry. *Anal. Chem.* **2009**, *81*, 10038−10048.

(9) Langman, L. J.; Kapur, B. M. Toxicology: Then and now. *Clin. Biochem.* **2006**, *39*, 498−510.

(10) Bethem, R.; Boison, J.; Gale, J.; Heller, D.; Lehotay, S.; Loo, J.; Musser, S.; Price, P.; Stein, S. Establishing the fitness for purpose of mass spectrometric methods. *J. Am. Soc. Mass. Spectrom.* **2003**, *14*, 528−541.

(11) McLafferty, F. W.; Stauffer, D. A.; Loh, S. Y.; Wesdemiotis, C. Unknown Identification Using Reference Mass Spectra. Quality Evaluation of Databases. *J. Am. Soc. Mass Spectrom.* **1999**, *10*, 1229−1240.

(12) Stein, S. E. Mass Spectral Reference Libraries: An Ever-Expanding Resource for Chemical Identification. *Anal. Chem.* **2012**, *84*, 7274−7282.

(13) Samokhin, A.; Sotnezova, K.; Lashin, V.; Revelsky, I. Evaluation of Mass Spectral Library Search Algorithms Implemented in Commercial Software. *J. Mass Spectrom.* **2015**, *50*, 820−825.

(14) Abrahamsson, S.; Haggstrom, G.; Stenhagen, E.An information retrieval system for organic mass spectrometry. In *14th Annual Conference on Mass Spectrometry and Allied Topics*, May 22−27, 1966, Dallas, TX; ASTM, 1966; pp 522−527.

(15) Mathews, R. J.; Morrison, J. D. Comparative Study of Methods of Computer-Matching Mass-Spectra. *Aust. J. Chem.* **1974**, *27*, 2167−2173.

(16) McLafferty, F. W.; Loh, S. Y.; Stauffer, D. B. Computer Identification of Mass Spectra. In *Computer-Enhanced Analytical Spectroscopy*; Springer, 1990; Vol 2, pp 163−181.

(17) Hertz, H. S.; Hites, R. A.; Biemann, K. Identification of Mass Spectra by Computer-Searching a File of Known Spectra. *Anal. Chem.* **1971**, *43*, 681−691.

(18) Rasmussen, G. T.; Isenhour, T. L. The Evaluation of Mass Spectral Search Algorithms. *J. Chem. Inf. Comput. Sci.* **1979**, *19*, 179−186.

(19) *ASTM D6420-18: Standard Test Method for Determination of Gaseous Organic Compounds by Direct Interface Gas Chromatography-Mass Spectrometry*; ASTM, 2018.

(20) *Association for Official Racing Chemists: Guidelines for the Minimum Criteria for Identification by Chromatography and Mass Spectrometry*; AORC, 2016.

(21) *Federal Bureau of Investigation: Guidelines for Comparison of Mass Spectra*; FBI, 2007.

(22) *Standard for Mass Spectral Data Acceptance for Definitive Identification*; NIST, 2014.

(23) *Science and Technology Pesticide Data Program: SOP*; United States Department of Agriculture Agricultural Marketing Service, 2017.

(24) *United Nations Office on Drugs and Crime: Guidance for the Validation of Analytical Methodology and Calibration of Equipment used for Testing of Illicit Drugs in Seized Materials and Biological Specimens*; United Nations, 2009.

(25) *WADA Technical Document: Minimum Criteria for Chromatographic-Mass Spectrometric Confirmation of the Identity of Analytes for Doping Control Purposes*; WADA, 2015.

(26) *European Commission Directorate-General for Health and Food Safety: Guidance Document on Analytical Quality Control and Method Validation Procedures for Pesticides Residues Analysis in Food and Feed*; European Commission, 2015.

(27) *FDA Food and and Veterinary Medicine Science and Research Steering Committee: Acceptance Criteria for Confirmation of Identity of Chemical Residues Using Exact Mass Data within the Office of Foods and Veterinary Medicine*; FDA, 2015.

(28) Davidson, J. T.; Lum, B. J.; Nano, G.; Jackson, G. P. Comparison of Measured and Recommended Acceptance Criteria for the Analysis of Seized Drugs using Gas Chromatography-Mass Spectrometry (GC-MS). *Forensic Chem.* **2018**, *10*, 15−26.
(29) Ausloos, P.; Clifton, C. L.; Lias, S. G.; Mikaya, A. I.; Stein, S. E.; Tchekhovskoi, D. V.; Sparkman, O. D.; Zaikin, V.; Zhu, D. The Critical Evaluation of a Comprehensive Mass Spectral Library. *J. Am. Soc. Mass Spectrom.* **1999**, *10*, 287−299.
(30) Kelly, K.; Bell, S. Evaluation of the reproducibility and repeatability of GCMS retention indices and mass spectra of novel psychoactive substances. *Forensic Chem.* **2018**, *7*, 10−18.
(31) Kelly, K.; Brooks, S.; Bell, S. The effect of mass spectrometry tuning frequency and criteria on ion relative abundances of cathinones and cannabinoids. *Forensic Chem.* **2019**, *12*, 58−65.
(32) Demuth, W.; Karlovits, M.; Varmuza, K. Spectral similarity versus structural similarity: mass spectrometry. *Anal. Chim. Acta* **2004**, *516*, 75−85.
(33) Samokhin, A. S.; Revel'skii, A. I.; Chepelyanskii, D. A.; Revel'skii, I. A. Possibility of the Reliable Identification of Unknown Compounds using an MS Search Program and a Commercial Electron Ionization Mass Spectral Database. *J. Anal. Chem.* **2011**, *66*, 1474−1476.
(34) Oberacher, H.; Pavlic, M.; Libiseller, K.; Schubert, B.; Sulyok, M.; Schuhmacher, R.; Csaszar, E.; Köfeler, H. C. On the Inter-Instrument and the Inter-Laboratory Transferability of a Tandem Mass Spectral Reference Library: 2. Optimization and Characterization of the Search Algorithm. *J. Mass Spectrom.* **2009**, *44*, 494−502.
(35) Oberacher, H.; Whitley, G.; Berger, B. Evaluation of the Sensitivity of the 'Wiley Registry of Tandem Mass Spectral Data, MSforID' with MS/MS Data of the 'NIST/NIH/EPA Mass Spectral Library'. *J. Mass Spectrom.* **2013**, *48*, 487−496.
(36) Oberacher, H.; Whitley, G.; Berger, B.; Weinmann, W. Testing an Alternative Search Algorithm for Compound Identification with the 'Wiley Registry of Tandem Mass Spectral Data, MSforID'. *J.Mass Spectrom.* **2013**, *48*, 497−504.
(37) Stein, S. E.; Scott, D. R. Optimization and Testing of Mass Spectral Library Search Algorithms for Compound Identification. *J. Am. Soc. Mass Spectrom.* **1994**, *5*, 859−866.
(38) Stein, S. E. Estimating Probabilities of Correct Identification From Results of Mass Spectral Library Searches. *J. Am. Soc. Mass Spectrom.* **1994**, *5*, 316−323.
(39) Hu, Q.; Zhang, J.; Chen, P.; Wang, B. Compound identification via deep classification model for electron-ionization mass spectrometry. *Int. J. Mass Spectrom.* **2021**, *463*, 116540.
(40) Koo, I.; Kim, S.; Zhang, X. Comparative Analysis of Mass Spectral Matching-Based Compound Identification in Gas Chromatography-Mass Spectrometry. *J. Chromatogr. A* **2013**, *1298*, 132−138.
(41) Kim, S.; Koo, I.; Wei, X.; Zhang, X. A Method of Finding Optimal Weight Factors for Compound Identification in Gas Chromatography-Mass Spectrometry. *Bioinformatics* **2012**, *28*, 1158−1163.
(42) McLafferty, F. W.; Zhang, M.-Y.; Stauffer, D. B.; Loh, S. Y. Comparison of Algorithms and Databases for Matching Unknown Mass Spectra. *J. Am. Soc. Mass Spectrom.* **1998**, *9*, 92−95.
(43) Domokos, L.; Henneberg, D.; Weimann, B. Optimization of Search Algorithms for a Mass-Spectra Library. *Anal. Chim. Acta* **1983**, *150*, 37−44.
(44) Domokos, L.; Hennberg, D.; Weimann, B. Computer-Aided Identification of Compounds by Comparison of Mass-Spectra. *Anal. Chim. Acta* **1984**, *165*, 61−74.
(45) Moorthy, A. S.; Sisco, E. The Min-Max Test: An Objective Method for Discriminating Mass Spectra. *Anal. Chem.* **2021**, *93*, 13319−13325.
(46) Wallace, W. E.; Ji, W. H.; Tchekhovskoi, D. V.; Phinney, K. W.; Stein, S. E. Mass Spectral Library Quality Assurance by Inter-Library Comparison. *J. Am. Soc. Mass. Spectrom.* **2017**, *28*, 733−738.
(47) Du, Y. M.; Hu, Y.; Xia, Y.; Ouyang, Z. Power Normalization for Mass Spectrometry Data Analysis and Analytical Method Assessment. *Anal. Chem.* **2016**, *88*, 3156−3163.
(48) Van Marlen, G.; Dijkstra, A.; Van't Klooster, H. A. Influence of Errors and Matching Criteria Upon the Retrieval of Binary Coded Low Resolution Mass-Spectra. *Anal. Chem.* **1979**, *51*, 420−423.
(49) Koo, I.; Kim, S.; Shi, B.; Lorkiewicz, P.; Song, M.; McClain, C.; Zhang, X. EIder: A Compound Identification Tool for Gas Chromatography Mass Spectrometry Data. *J. Chromatogr. A* **2016**, *1448*, 107−114.
(50) Kim, S.; Koo, I.; Jeong, J.; Wu, S.; Shi, X.; Zhang, X. Compound Identification using Partial and Semipartial Correlations for Gas Chromatography-Mass Spectrometry Data. *Anal. Chem.* **2012**, *84*, 6477−6487.
(51) Koo, I.; Zhang, X.; Kim, S. Wavelet- and Fourier-Transform-Based Spectrum Similarity Approaches to Compound Identification in Gas Chromatography/Mass Spectrometry. *Anal. Chem.* **2011**, *83*, 5631−5638.
(52) Wei, X.; Koo, I.; Kim, S.; Zhang, X. Compound Identification in GC-MS by Simultaneously Evaluating Mass Spectrum and Retention Index. *Analyst* **2014**, *139*, 2507−2514.
(53) Koo, I.; Shi, X.; Kim, S.; Zhang, X. iMatch2: Compound Identification using Retention Index for Analysis of Gas Chromatography-Mass Spectrometry Data. *J. Chromatogr. A* **2014**, *1337*, 202−210.
(54) Jeong, J.; Shi, X.; Zhang, X.; Kim, S.; Shen, C. An Empirical Bayes Model using a Competition Score for Metabolite Identification in Gas Chromatography Mass Spectrometry. *BMC Bioinformatics* **2011**, *12*, 392.
(55) Stein, S. E. An Integrated Method for Spectrum Extraction and Compound Identification from Gas Chromatography/Mass Spectrometry Data. *J. Am. Soc. Mass Spectrom.* **1999**, *10*, 770−781.
(56) Matsuo, T.; Tsugawa, H.; Miyagawa, H.; Fukusaki, E. Integrated Strategy for Unknown EI-MS Identification Using Quality Control Calibration Curve, Multivariate Analysis, EI-MS Spectral Database, and Retention Index Prediction. *Anal. Chem.* **2017**, *89*, 6766−6773.
(57) Schymanski, E. L.; Gallampois, C. M.; Krauss, M.; Meringer, M.; Neumann, S.; Schulze, T.; Wolf, S.; Brack, W. Consensus Structure Elucidation Combining GC/EI-MS, Structure Generation, and Calculated Properties. *Anal. Chem.* **2012**, *84*, 3287−3295.
(58) Smith, D. H.; Achenbach, M.; Yeager, W. J.; Anderson, P. J.; Fitch, W. L.; Rindfleisch, T. C. Quantitative Comparison of Combined Gas Chromatographic-Mass Spectrometric Profiles of Complex-Mixtures. *Anal. Chem.* **1977**, *49*, 1623−1632.
(59) Sisco, E.; Burns, A.; Moorthy, A. S. A Framework for the Development of Targeted Gas Chromatography Mass Spectrometry (GC-MS) Methods: Synthetic Cannabinoids. *J. Forensic Sci.* **2021**, *66*, 1908−1918.
(60) Dromey, R. G.; Stefik, M. J.; Rindfleisch, T. C.; Duffield, A. M. Extraction of Mass-Spectra Free of Background and Neighboring Component Contributions from Gas Chromatography Mass Spectrometry Data. *Anal. Chem.* **1976**, *48*, 1368−1375.
(61) Samokhin, A. S.; Revel'skii, I. A. Application of Principal Component Analysis to the Extraction of Pure Mass Spectra in Chemical Analysis by Gas Chromatography/Mass Spectrometry. *J. Anal. Chem.* **2010**, *65*, 1481−1488.
(62) Julian, R. K.; Higgs, R. E.; Gygi, J. D.; Hilton, M. D. A Method for Quantitatively Differentiating Crude Natural Extracts using High-Performance Liquid Chromatography Electrospray Mass Spectrometry. *Anal. Chem.* **1998**, *70*, 3249−3254.
(63) Grotch, S. L. Automatic Identification of Chemical Spectra - Goodness of Fit Measure Derived from Hypothesis Testing. *Anal. Chem.* **1975**, *47*, 1285−1289.
(64) Hites, R. A.; Biemann, K. A Computer-Compatible Digital Data Acquisition System for Fast-Scanning Single-Focusing Mass Spectrometers. *Anal. Chem.* **1967**, *39*, 965−970.
(65) Meyerson, S. Derivation of Mass Spectra of Individual Compounds from the Spectra of Mixtures. *Anal. Chem.* **1959**, *31*, 174−175.
(66) Domokos, L.; Henneberg, D. A Correlation Method in Library Search. *Anal. Chim. Acta* **1984**, *165*, 75−86.

https://doi.org/10.1021/jasms.3c00089
J. Am. Soc. Mass Spectrom. 2023, 34, 1248−1262

(67) Sobcov, H. Analysis of Liquid Hydrocarbon Mixtures by Mass Spectrometry - Application of Matrices and IBM Techniques. *Anal. Chem.* **1952**, *24*, 1386−1388.
(68) Halket, J. M. K.; Reed, R. I. Analysis of Mixtures.5. *Org. Mass Spectrom.* **1976**, *11*, 881−887.
(69) Satpathy, G.; Tyagi, Y. K.; Gupta, R. K. A Novel Optimised and Validated Method for Analysis of Multi-Residues of Pesticides in Fruits and Vegetables by Microwave-Assisted Extraction (MAE)-Dispersive Solid-Phase Extraction (D-SPE)-Retention Time Locked (RTL)-Gas Chromatography-Mass Spectrometry with Deconvolution Reporting Software (DRS). *Food Chem.* **2011**, *127*, 1300−1308.
(70) You, Y.; Song, L.; Young, M. D.; van der Wielen, M.; Evans-Nguyen, T.; Riedel, J.; Shelley, J. T. Unsupervised Reconstruction of Analyte-Specific Mass Spectra Based on Time-Domain Morphology with a Modified Cross-Correlation Approach. *Anal. Chem.* **2021**, *93*, 5009−5014.
(71) Drablos, F. Symmetric Distance Measures for Mass-Spectra. *Anal. Chim. Acta* **1987**, *201*, 225−239.
(72) Bodnar Willard, M. A.; McGuffin, V. L.; Smith, R. W. Statistical Comparison of Mass Spectra for Identification of Amphetamine-Type Stimulants. *Forens. Sci. Int.* **2017**, *270*, 111−120.
(73) Bodnar Willard, M. A.; Waddell Smith, R.; McGuffin, V. L. Statistical Approach to Establish Equivalence of Unabbreviated Mass Spectra. *Rapid Commun. Mass Spectrom.* **2014**, *28*, 83−95.
(74) Stuhmer, E. L.; McGuffin, V. L.; Waddell Smith, R. Discrimination of seized drug positional isomers based on statistical comparison of electron-ionization mass spectra. *Forensic Chem.* **2020**, *20*, 100261.
(75) Crawford, L. R.; Morrison, J. D. Computer Methods in Analytical Mass Spectrometry: Identification of an Unknown Compound in a Catalog. *Anal. Chem.* **1968**, *40*, 1464−1469.
(76) McLafferty, F. W.; Hertel, R. H.; Villwock, R. D. Probability Based Matching of Mass Spectra. Rapid Identification of Specific Compounds in Mixtures. *J. Mass. Spectrom.* **1974**, *9*, 690−702.
(77) Bafna, V.; Edwards, N. SCOPE: A Probabilistic Model for Scoring Tandem Mass Spectra against a Peptide Database. *Bioinformatics* **2001**, *17*, S13−S21.
(78) Gatto, L.; Hansen, K. D.; Hoopmann, M. R.; Hermjakob, H.; Kohlbacher, O.; Beyer, A. Testing and Validation of Computational Methods for Mass Spectrometry. *J. Proteome Res.* **2016**, *15*, 809−814.
(79) Park, C. Y.; Klammer, A. A.; KäLl, L.; Maccoss, M. J.; Noble, W. S. Rapid and Accurate Peptide Identification from Tandem Mass Spectra. *J. Proteome Res.* **2008**, *7*, 3022−3027.
(80) Wan, K. X.; Vidavsky, I.; Gross, M. L. Comparing Similar Spectra: From Similarity Index to Spectral Contrast Angle. *J. Am. Soc. Mass. Spectrom.* **2002**, *13*, 85−88.
(81) Yang, X.; Neta, P.; Stein, S. E. Extending a Tandem Mass Spectral Library to Include MS(2) Spectra of Fragment Ions Produced In-Source and MS(n) Spectra. *J. Am. Soc. Mass Spectrom.* **2017**, *28*, 2280−2287.
(82) Fu, Y.; Yang, Q.; Sun, R.; Li, D.; Zeng, R.; Ling, C. X.; Gao, W. Exploiting the kernel trick to correlate fragment ions for peptide identification via tandem mass spectrometry. *Bioinformatics* **2004**, *20*, 1948−1954.
(83) Driver, T.; Averbukh, V.; Frasinski, L. J.; Marangos, J. P.; Edelson-Averbukh, M. Two-Dimensional Partial Covariance Mass Spectrometry for the Top-Down Analysis of Intact Proteins. *Anal. Chem.* **2021**, *93*, 10779−10788.
(84) Driver, T.; Bachhawat, N.; Frasinski, L. J.; Marangos, J. P.; Averbukh, V.; Edelson-Averbukh, M. Chimera Spectrum Diagnostics for Peptides Using Two-Dimensional Partial Covariance Mass Spectrometry. *Molecules* **2021**, *26*, 3728.
(85) Driver, T.; Bachhawat, N.; Pipkorn, R.; Frasiński, L. J.; Marangos, J. P.; Edelson-Averbukh, M.; Averbukh, V. Proteomic Database Search Engine for Two-Dimensional Partial Covariance Mass Spectrometry. *Anal. Chem.* **2021**, *93*, 14946−14954.
(86) Driver, T.; Cooper, B.; Ayers, R.; Pipkorn, R.; Patchkovskii, S.; Averbukh, V.; Klug, D. R.; Marangos, J. P.; Frasinski, L. J.; Edelson-Averbukh, M. Two-Dimensional Partial-Covariance Mass Spectrometry of Large Molecules Based on Fragment Correlations. *Physical Review X* **2020**, *10*, 041004.
(87) Dromey, R. G. Optimum scaling of mass spectra for computer-matching. *Anal. Chim. Acta* **1979**, *112*, 133−141.
(88) Atwater, B. L.; Stauffer, D. B.; McLafferty, F. W.; Peterson, D. W. Reliability Ranking and Scaling Improvements to the Probability Based Matching System for Unknown Mass Spectra. *Anal. Chem.* **1985**, *57*, 899−903.
(89) Sigman, M. E.; Clark, C. D. Two-dimensional correlation spectroscopy techniques applied to ion trap tandem mass spectrometric analysis: nitroaromatics. *Rapid Commun. Mass Spectrom.* **2005**, *19*, 3731−3736.
(90) Kornig, S.; Hoogerbrugge, R.; Vanwitzenburg, W. R.; Kistemaker, P. G. Tandem Mass-Spectrometry Resolution Enhancement by Multivariate-Analysis. *Int. J. Mass Spectrom. Ion Processes* **1989**, *89*, 111−124.
(91) Garrido, B.; Cavalcanti, G.; Leal, F.; Aguiar, P.; Padilha, M.; Radler de Aquino Neto, F. Principal component analysis (PCA) of the fragmentation patterns of anabolic steroids by tandem mass spectrometry with electrospray ionization. *Braz. J. Anal. Chem.* **2012**, *2*, 309−315.
(92) Wold, S.; Christie, O. H. J. Extraction of Mass Spectral Information by a Combination of Autocorrelation and Principal Components Models. *Anal. Chim. Acta* **1984**, *165*, 51−59.
(93) Varmuza, K.; Werther, W.; Henneberg, D.; Weimann, B. Computer-Aided Interpretation of Mass-Spectra by a Combination of Library Search with Principal Component Analysis. *Rapid Commun. Mass Spectrom.* **1990**, *4*, 159−162.
(94) Harris, D. N.; Hokanson, S.; Miller, V.; Jackson, G. P. Fragmentation Differences in the EI Spectra of Three Synthetic Cannabinoid Positional Isomers: JWH-250, JWH-302, and JWH-201. *Int. J. Mass Spectrom.* **2014**, *368*, 23−29.
(95) Gilbert, N.; Mewis, R. E.; Sutcliffe, O. B. Classification of Fentanyl Analogues through Principal Component Analysis (PCA) and Hierarchical Clustering of GC-MS Data. *Forensic Chem.* **2020**, *21*, 100287.
(96) Hill, H. C.; Reed, R. I.; Robert-Lopes, M. T. Mass Spectra and Molecular Structure 1: Correlation Studies and Metastable Transitions. *J. Chem. Soc. C* **1968**, 93−101.
(97) Sotnezova, K. M.; Samokhin, A. S.; Revelsky, I. A. Use of PLS Discriminant Analysis for Revealing the Absence of a Compound in an Electron Ionization Mass Spectral Database. *J. Anal. Chem.* **2017**, *72*, 1419−1425.
(98) Setser, A. L.; Waddell Smith, R. Comparison of Variable Selection Methods Prior to Linear Discriminant Analysis Classification of Synthetic Phenethylamines and Tryptamines. *Forensic Chem.* **2018**, *11*, 77−86.
(99) Davidson, J. T.; Jackson, G. P. The differentiation of 2,5-dimethoxy-N-(N-methoxybenzyl)phenethylamine (NBOMe) isomers using GC retention indices and multivariate analysis of ion abundances in electron ionization mass spectra. *Forensic Chem.* **2019**, *14*, 100160.
(100) Bonetti, J. Mass Spectral Differentiation of Positional Isomers using Multivariate Statistics. *Forensic Chem.* **2018**, *9*, 50−61.
(101) Liliedahl, R. E.; Davidson, J. T. The Differentiation of Synthetic Cathinone Isomers using GC-EI-MS and Multivariate Analysis. *Forensic Chem.* **2021**, *26*, 100349.
(102) Rotter, H.; Varmuza, K. Computer-Aided Interpretation of Steroid Mass-Spectra by Pattern-Recognition Methods. 3: Computation of Binary Classifiers by Linear-Regression. *Anal. Chim. Acta Comp. Technol. Opt.* **1978**, *103*, 61−71.
(103) Rotter, H.; Varmuza, K. Computer-Aided Interpretation of Steroid Mass-Spectra by Pattern-Recognition Methods. 2: Influence of Mass-Spectral Preprocessing on Classification by Distance Measurement to Centers of Gravity. *Anal. Chim. Acta* **1977**, *95*, 25−32.
(104) Harrington, P. d. B.; Voorhees, K. J. Multivariate Rule Building Expert System. *Anal. Chem.* **1990**, *62*, 729−734.
(105) Harrington, P. d. B.; Street, T. E.; Voorhees, K. J.; Radicati di Brozolo, F.; Odom, R. W. A Rule-Building Expert System for Classification of Mass Spectra. *Anal. Chem.* **1989**, *61*, 715−719.

https://doi.org/10.1021/jasms.3c00089
J. Am. Soc. Mass Spectrom. 2023, 34, 1248−1262

(106) Samokhin, A.; Revelsky, I. Distinguishing by Principal Component Analysis o-Xylene, m-Xylene, p-Xylene and Ethylbenzene using Electron Ionization Mass Spectrometry. *Eur. J. Mass Spectrom.* **2011**, *17*, 477−480.
(107) Bonetti, J. L.; Samanipour, S.; van Asten, A. C. Utilization of Machine Learning for the Differentiation of Positional NPS Isomers with Direct Analysis in Real Time Mass Spectrometry. *Anal. Chem.* **2022**, *94*, 5029−5040.
(108) Crawford, L. R.; Morrison, J. D. Computer Methods in Analytical Mass Spectrometry: Development of Programs for Analysis of Low Resolution Mass Spectra. *Anal. Chem.* **1971**, *43*, 1790−1795.
(109) Zhang, J.; Wei, X.-L.; Zheng, C.-H.; Wang, B.; Wang, F.; Chen, P. Compound Identification using Random Projection for Gas Chromatography-Mass Spectrometry Data. *Int. J. Mass Spectrom.* **2016**, *407*, 16−21.
(110) Dührkop, K.; Fleischauer, M.; Ludwig, M.; Aksenov, A. A.; Melnik, A. V.; Meusel, M.; Dorrestein, P. C.; Rousu, J.; Böcker, S. SIRIUS 4: a Rapid Tool for Turning Tandem Mass Spectra into Metabolite Structure Information. *Nat. Methods* **2019**, *16*, 299−302.
(111) Koshute, P.; Hagan, N.; Jameson, N. J. Machine Learning Model for Detecting Fentanyl Analogs from Mass Spectra. *Forensic Chem.* **2022**, *27*, 100379.
(112) Anderson, D. C.; Li, W.; Payan, D. G.; Noble, W. S. A New Algorithm for the Evaluation of Shotgun Peptide Sequencing in Proteomics: Support Vector Machine Classification of Peptide MS/MS Spectra and SEQUEST Scores. *J. Proteome Res.* **2003**, *2*, 137−146.
(113) Käll, L.; Canterbury, J. D.; Weston, J.; Noble, W. S.; MacCoss, M. J. Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nat. Methods* **2007**, *4*, 923−925.
(114) Elias, J. E.; Gibbons, F. D.; King, O. D.; Roth, F. P.; Gygi, S. P. Intensity-Based Protein Identification by Machine Learning from a Library of Tandem Mass Spectra. *Nat. Biotechnol.* **2004**, *22*, 214−219.
(115) Zhu, R. L.; Jonas, E. Rapid Approximate Subset-Based Spectra Prediction for Electron Ionization-Mass Spectrometry. *Anal. Chem.* **2023**, *95*, 2653−2663.
(116) Allen, A.; Cooper, D.; Kiser, W.; Cottreli, R. The Cocaine Diastereoisomers. *J. Forensic Sci.* **1981**, *26*, 12−26.
(117) Smith, R. The Mass Spectrum of Cocaine. *J. Forensic Sci.* **1997**, *42*, 475−480.
(118) Smith, R. M.; Casale, J. F. The Mass Spectrum of Cocaine: Deuterium Labeling and MS/MS Studies. *Microgram J.* **2010**, *7*, 16−41.
(119) Tal'roze, V. L.; Tantsyrev, G. D.; Raznikov, V. V. Minimum of Information Sufficient to Identify Individual Organic Substances by Concidence of Their Mass-Spectrum Lines. *Dokl Akad Nauk SSSR* **1964**, *159*, 182−185.
(120) Raznikov, V. V.; Tal'roze, V. L. Teaching an Electron Computer of Recognizing 2 Classes of Organic Compounds from Several Like Mass-Spectrum Lines. *Dokl Akad Nauk SSSR* **1966**, *170*, 379−382.
(121) Olson, M. T.; Blank, P. S.; Sackett, D. L.; Yergey, A. L. Evaluating Reproducibility and Similarity of Mass and Intensity Data in Complex Spectra: Applications to Tubulin. *J. Am. Soc. Mass Spectrom.* **2008**, *19*, 367−374.
(122) Lasko, T. A.; Bhagwat, J. G.; Zou, K. H.; Ohno-Machado, L. The Use of Receiver Operating Characteristic Curves in Biomedical Informatics. *J. Biomed. Inform.* **2005**, *38*, 404−415.
(123) McLafferty, F. W. Performance Prediction and Evaluation of Systems for Computer Identification of Spectra. *Anal. Chem.* **1977**, *49*, 1441−1443.
(124) SWGDRUG MS Library Version 3.11. https://www.swgdrug.org/ (accessed june 1, 2022).
(125) Nishimura, T. Unimolecular Dissociations in Mass Spectrometry. In *Fundamentals of Mass Spectrometry*; Hiraoka, K., Ed.; Springer: New York, 2013.
(126) Sleno, L.; Volmer, D. Ion Activation Methods for Tandem Mass Spectrometry. *J. Mass Spectrom.* **2004**, *39*, 1091−1112.
(127) Rosenstock, H. M.; Wallenstein, M. B.; Wahrhaftig, A. L.; Eyring, H. Absolute Rate Theory for Isolated Systems and the Mass Spectra of Polyatomic Molecules. *Proc. Natl. Acad. Sci. U.S.A.* **1952**, *38*, 667.
(128) Samokhin, A. Spectral Skewing in Gas Chromatography-Mass Spectrometry: Misconceptions and Realities. *J. Chrom. A* **2018**, *1576*, 113−119.
(129) Bayat, P.; Lesage, D.; Cole, R. B. Tutorial: Ion Activation in Tandem Mass Spectrometry Using Ultra-High Resolution Instrumentation. *Mass Spectrom. Rev.* **2020**, *39*, 680−702.
(130) Kassel, L. S. Studies in Homogeneous Gas Reactions II: Introduction of Quantum Theory. *J. Phys. Chem.* **1928**, *32*, 1065−1079.
(131) Marcus, R. A. Unimolecular Dissociations and Free Radical Recombination Reactions. *J. Chem. Phys.* **1952**, *20*, 359−364.
(132) Rice, O. K.; Ramsperger, H. C. Theories of Unimolecular Gas Reactions at Low Pressures I. *J. Am. Chem. Soc.* **1927**, *49*, 1617−1629.
(133) Rice, O. K.; Ramsperger, H. C. Theories of Unimolecular Gas Reactions at Low Pressures II. *J. Am. Chem. Soc.* **1928**, *50*, 617−620.
(134) Baer, T.; Mayer, P. M. Statistical Rice-Ramsperger-Kassel-Marcus Quasiequilibrium Theory Calculations in Mass Spectrometry. *J. Am. Soc. Mass Spectrom.* **1997**, *8*, 103−115.
(135) Vekey, K. Internal Energy Effects in Mass Spectrometry. *J. Mass Spectrom.* **1996**, *31*, 445−463.
(136) Bauer, C. A.; Grimme, S. How to Compute Electron Ionization Mass Spectra from First Principles. *J. Phys. Chem. A* **2016**, *120*, 3755−3766.
(137) Kassel, L. S. Studies in Homogeneous Gas Reactions. I. *J. Phys. Chem.* **1928**, *32*, 225−242.
(138) Friedman, L.; Long, F. A.; Wolfsberg, M. Ionization Efficiency Curves and the Statistical Theory of Mass Spectra. *J. Chem. Phys.* **1957**, *26*, 714−715.
(139) Shukla, A. K.; Futrell, J. H. Tandem Mass Spectrometry: Dissociation of Ions by Collisional Activation. *J. Mass Spectrom.* **2000**, *35*, 1069−1090.
(140) Garcia, B. A. What Does the Future Hold for Top Down Mass Spectrometry? *J. Am. Soc. Mass. Spectrom.* **2010**, *21*, 193−202.
(141) Greisch, J.-F.; den Boer, M. A.; Lai, S.-H.; Gallagher, K.; Bondt, A.; Commandeur, J.; Heck, A. J. R. Extending Native Top-Down Electron Capture Dissociation to MDa Immunoglobulin Complexes Provides Useful Sequence Tags Covering Their Critical Variable Complementarity-Determining Regions. *Anal. Chem.* **2021**, *93*, 16068−16075.
(142) Morrison, J. D.; Crawford, L. R. Computer Methods in Analytical Mass Spectrometry. *Anal. Chem.* **1968**, *40*, 1464−1469.
(143) Galton, F. Regression Towards Mediocrity in Hereditary Stature. *J. Anthr. Inst. G. B. Ire.* **1886**, *15*, 246−263.
(144) Wise, B.; Ricker, N. L. Recent Advances in Multivariate Statistical Process Control: Improving Robustness and Sensitivity. In *IFAC Advanced Control of Chemical Processes*; Permagon: Tolouse, France, 1991; pp 125−130.
(145) Jalali-Heravi, M.; Ebrahimi-Najafabadi, H.; Khodabandehloo, A. Use of Kernel Orthogonal Projection to Latent Structure in Modeling of Retention Indices of Pesticides. *Qsar & Comb. Sci.* **2009**, *28*, 1432−1441.

https://doi.org/10.1021/jasms.3c00089
J. Am. Soc. Mass Spectrom. 2023, 34, 1248−1262