2023

The Influence of Instrumental Sources of Variance on Mass Spectral Comparison Algorithms

Isabel Cristina Galvez Valencia
icg00001@mix.wvu.edu

The Influence of Instrumental Sources of Variance on Mass Spectral Comparison Algorithms

Isabel Cristina Gálvez Valencia

Thesis submitted to the Eberly College of Arts and Sciences at West Virginia University
in partial fulfillment of the requirements for the degree of
Master of Science in
Forensic and Investigative Science

Glen P. Jackson, Ph.D., (Chair)

Jacqueline Speir, Ph.D.

Stephen Valentine, Ph.D.

Department of Forensic and Investigative Science

Morgantown, West Virginia
2023

Keywords: General linear model, Mass spectrometry, Multivariate analysis, Algorithm

Abstract

The Influence of Instrumental Sources of Variance on Mass Spectral Comparison Algorithms

Isabel Cristina Gálvez Valencia

Current search algorithms for the identification of substances based only on their electron ionization mass spectra provide the correct compound as their top result approximately 80% of the time. One contributing factor to the ~20% deviation in the first-hit recognition rate is that traditional algorithms work by comparing the unknown spectrum to an 'ideal' or consensus spectrum of each reference compound. The inclusion of replicate reference spectra in a database has been shown to improve the probability of ranking the correct identity in the number one position, but the variance in ion abundances caused by different conditions or different instruments remains an intractable problem and the major source of uncertainty in mass spectral identification.

To assess the relative contributions of different factors to the spectral variance of replicate spectra, this study initially considered the repeller voltage, focus lens voltage, and ion energy as primary parameters. A three-factor, three-level, full-factorial design of experiments was conducted using cocaine as a model compound. A library of cocaine spectra was collected with a gas chromatography-electron ionization-mass spectrometer (GC-EI-MS) by extracting each spectrum across the eluting peak. The 20 most abundant ions in the library of cocaine spectra were extracted to assess the contribution of each instrument parameter on the variance in ion abundances by performing multivariate analysis of variance (MANOVA). Results showed that these instrument parameters were responsible for only ~3% of the total variance in the normalized abundances. This initial finding prompted a subsequent study that monitored the branching ratios of cocaine during random fluctuations in the vacuum chamber pressure. Random changes in vacuum pressure accounted for ~90% of the natural variance in the relative ion abundances of the two most abundant peaks of cocaine (not including the base peak).

The database of 389 cocaine spectra was then used to compare the traditional consensus approaches to spectral matching with two variants of a novel algorithm called the Expert Algorithm for Substance Identification (EASI). EASI uses multivariate linear modeling to predict the ion abundances of 20 ions in each spectrum, assuming that each of the 20 ion abundances is continuously dependent on the other 19 ion abundances. One variant of this model includes intercepts in the linear models, and the other does not. To assess the effect of spectral variance on spectral identifications, traditional measures of spectral similarity or dissimilarity were calculated between each query spectrum and the consensus cocaine spectrum, including the Pearson product-moment correlation (PPMC) coefficients, mean absolute residuals (MARs), Euclidean distances, and NIST scores. These metrics were then used as binary classifiers to obtain true positives, true negatives, false positives, and false negatives at a range of decision thresholds. The models were tested on a database of spectra that included more than 300 cocaine spectra from different

laboratories, more than 700 spectra of 5 common drugs, and 10 spectra of cocaine diastereomers: allococaine, pseudococaine, and pseudoallococaine. The EASI models outperformed the consensus approach on every metric. EASI coupled with the PPMC values, MARs and Euclidean distances had accuracies greater than 90% with zero false positives, including spectra of cocaine diastereomers and cocaine collected on different instruments. The Mahalanobis distances to the training set as a binary classifier were also reported, and they were found to be as good or better than EASI at discriminating between cocaine and non-cocaine spectra.

Each measure of spectral similarity was used to build receiver operating characteristic (ROC) curves and calculate the area under the ROC curve (AUC). When taking only the cocaine diastereomers as known negatives, the EASI without a constant had the highest area under the curve (AUC=0.925), followed by EASI including a constant (AUC=0.907), and lastly the consensus model with (AUC=0.829). This work shows that random variations in vacuum pressure are responsible for most of the short-term variance in replicate mass spectra and that a model (EASI) that accounts for cross-correlations between the different fragment ions allow superior compound identification to traditional algorithms.

Table of Contents

List of abbreviations

ANOVA	Analysis of Variance
AUC	Area Under the Curve
DFT	Discrete Fourier Transform
DWT	Discrete Wavelet Transform
EASI	Expert Algorithm for Substance Identification
EASI WO	Expert Algorithm for Substance Identification Without a constant
EI	Electron Ionization
EME	Ecgonine Methyl Ester
FN	False Negative
FNR	False Negative Rate
FP	False Positive
FPR	False Positive Rate
GC	Gas Chromatography
GLM	General Liner Model
HSS	Hybrid Similarity Search
KN	Known Negative
KP	Known Positive
LDA	Linear Discriminant Analysis
MANOVA	Multivariate Analysis of Variance
MAR	Mean Absolute Residual
MF	Match Factor
MS	Mass Spectrometer
NIST	National Institute of Standards and Technology
PBM	Probability-Based Matching
PCA	Principal Component Analysis
PFTBA	Perfluorotributylamine
PPMC	Pearson Product-Moment Correlation
QET	Quasi-Equilibrium Theory
RMP	Random Match Probability
ROC	Receiver Operating Characteristic
RRKM	Rice–Ramsperger–Kassel–Marcus
SPSS	Statistical Package for the Social Sciences
SSS	Simple Similarity Search
SWGDRUG	Scientific Working Group for the Analysis of Seized Drugs
TN	True Negative
TNR	True Negative Rate
TP	True Positive
TPR	True Positive Rate

List of tables

List of figures

1. Statement of problem

One of the main tasks of a forensic chemist is to identify controlled substances in seized drugs. This operation is usually accomplished using gas chromatography-electron ionization-mass spectrometry (GC-EI-MS), which produces mass spectra that can be compared to spectra of known or reference compounds through the help of search algorithms. However, current algorithms do not account for inter-instrument variance, so laboratories are required to perform spectral matching with reference standards collected on the same instrument, on the same day, under the same conditions as the query spectrum, and algorithms generally are not as successful at comparing results obtained on different instruments or in different laboratories. This presents a problem, especially for novel psychoactive substances (NPS), because reference materials can be expensive, hard to obtain, and hazardous to handle. If an algorithm could be developed that could tolerate the spectral variance caused by instrument variance, then labs could operate more safely, more quickly and with reduced costs, while improving the confidence in their drug identifications.

This project aims to develop and test a new mass spectral comparison algorithm for the identification of seized drugs. The project will first determine the effects of different parameters—such as the ion energy, repeller voltage and focus lens voltage—on replicate EI-MS spectra of cocaine and then test an algorithm that can identify cocaine, with improved confidence, from a wide array of seized drugs, including the three major diastereomers of cocaine.

2. Introduction

GC-EI-MS is the most widely used method for compound identification, including drugs of abuse.[1–5] This technique is classified by the Scientific Working Group for the Analysis of Seized Drugs (SWGDRUG) as a category A method, which means that it provides structural information with a level of selectivity that is among the highest of all analytical techniques.[6]

During electron ionization in the ion source of a mass spectrometer, molecular ions are produced via vertical (Franck-Condon) transitions that result in internal energies that can exceed the bond dissociation energy by tens of electron volts. Statistical theories, including Rice–Ramsperger–Kassel–Marcus (RRKM) and quasi-equilibrium theory (QET) describe how the excess internal energy is effectively distributed throughout the molecule before the molecule takes the time to move through a particular transition state to the fragments. The kinetics of unimolecular fragmentation through the various pathways depend on factors like the internal degrees of freedom of the ion, the internal (excitation) energy of the ion, the steric or entropic requirements of each transition state and the bond dissociation energies of different transition states.[7]

Once ions are created, a bias voltage of 10-40 V applied to the repeller electrode pushes the ions out of the ionization chamber towards a series of ion lenses. These lenses accelerate and focus the ion beam just before they enter the mass analyzer, which is typically a quadrupole mass analyzer. The quadrupole separates the ions by applying an appropriate ratio of rf and dc potentials to alter the stable trajectories of ions according to their mass-to-charge ratios (*m/z*).[7,8] Depending on the application, analysts then either interpret the fragmentation pattern according to common rules about fragmentation mechanisms or compare the measured peak intensities with those of reference spectra.

One of the earliest ways to identify an unknown compound was to directly compare mass peaks against a 'master deck' of standards that were previously sorted by their top ten highest peaks using an IBM computer.[9] Later, Crawford and Morrison tried different normalization methods for the spectra, comparing them through a discrepancy factor derived from the normalization equation. This method only considered the ion abundances as variables, which revealed other factors that had effects on the spectra, like impurities and differences between

instruments over peak heights.[10] A variation of this method was later used in 1971 to identify drugs. This method involved arranging both the queried and 'master deck' spectra by their top five most abundant peaks and matching their *m/z* values, if no match was found, the computer would look for a match within the top four ions, and so on, until there was only one peak left.[11]

Later, with the increased availability of computers, it became apparent that spectral interpretation had potential for automation and more complex statistical analyses. Grotch developed a new approach that proved to be suitable for digital computation.[12] The new algorithm consisted of comparing an unknown spectrum to a library of previously studied compounds. To ease the calculations, he encoded the data to one bit. Then, he compared the encoded peak value at each *m/z* to a known spectrum in a pairwise manner and assigned a value of 1 for each disagreement and 0 for each agreement. Comparisons with the lowest scores were considered the best matches. Here, the paradigm shifted from subjective human judgments to objective mathematical justifications for matching spectra.[12]

Knock and coworkers continued this research by using non-one-bit encoded data. This project revealed that different operating conditions on mass spectrometers could affect the breakdown pathways, thus influencing the mass spectra. They also introduced the importance of considering the peaks' intensity as a deciding factor when matching spectra.[13]

Following previous work, Hertz et al. introduced a measure of similarity that provides a quantitative result and considers the entire spectrum, not just the top five or ten most intense peaks.[14] The product of this research was the 'similarity index' (Eqn. 1), a ratio that is equivalent to the probability of agreement, which ranges from 0 for complete disagreement to 1 for complete agreement.

$$Similarity\ index = \frac{average\ weighted\ ratio}{fraction\ of\ unmatched\ intensities + 1} \qquad \text{Eqn. 1}$$

The similarity index is the average weighted ratio of the reference to an unknown reduced spectrum taken mass for mass.[14] The ratios are weighted by a specific factor depending on their normalized abundance. This weighing process allows the larger intensity peaks to be more significant than the smaller ones.[14] A similar algorithm is still used by NIST today.[15–17]

Next came the probability-based matching (PBM) technique by McLafferty and coworkers. PBM was an improvement because it considered how unique and how abundant the peaks were, as well as the absence of other peaks. This technique analyzed the probability that a specific compound was present in a sample by establishing a 'confidence index', *K* (Eqn. 2), which is defined as the summation of four individual probabilities *U*, *A*, *D* and *W*.

$$K = \sum (U_j - A_j - D + W_j) \qquad \text{Eqn. 2}$$

In Eqn. 2, U_j represents the probability that the abundance of the *j*th peak is greater than 50% of the base peak of a randomly selected spectrum, A_j is a modification on U_j in which A_j defines a minimum value for a particular mass abundance based on the relative abundance of the peak in the reference spectrum of the target compound. These two values had to be modified to account for different abundance distributions. The term D factors in the dilution of the sample, further correcting the abundance (D = 0 in a pure sample). W_j is the window tolerance and refers to the expected degree a sample has of matching the abundance requirements by chance, and it reflects experimental factors like reproducibility and background so that it will depend on mass and instrumentation.

A high *K* value indicates higher confidence in the identification of the unknown spectrum, whereas a low *K* value denotes the opposite. Results also show that *K* values are influenced by the sample size, where pure samples of average size (10 μg) can yield *K* values between 75 to 125, and a smaller sample of 0.1 μg will result in *K* values between 20 and 40.[1]

In the following years, as the PBM algorithm was popularized for illicit drug identification, several research groups contributed to improving the technique, which is still used by some Agilent instruments.[3,18–22] One of the most important developments was the peak flagging process. This action consisted of flagging the least abundant peaks not to be included in the final calculations.[23]

Parallel to the development of PBM algorithm, Atwater and coworkers developed a matching indicator, called the reliability value (RL), based on the predicted match reliability of the PBM. This matching indicator focused on providing the overall probability of a correct match instead of the probability that the unknown compound was present in the reference library, like previous metrics.[1,14] To accomplish this goal, the RL used the confidence value *K* (Eqn. 2), a quadratic scaling factor, the number of peak flagging operations, and whether the molecular ion was used for the matching process or not.[21] The scaling process used for the RL was based on the quadratic polynomial adjustment of the peak abundances of the reference spectrum to minimize the sum of squares between the queried and reference spectra.[21] This quadratic scaling process was added to compensate for variations in sample concentrations during scans or variations in mass abundance due to mass discrimination by quadrupole mass filters.[16]

Alternatively, other algorithms during the ‘90s were based on the statistical comparison of the unknown and the database spectra. The most popular methods include Euclidean distance,[16,24] absolute value distance,[16,24] and weighted or unweighted dot-products, which uses the cosine of the angle between unknown and reference vector representations.[16,24] In the dot-product comparison, a value of 1 conveys perfect correlation, and therefore higher confidence in the identification, and a value of 0 denotes the opposite.[16]

To compare the performance of the main algorithms at the time, Stein and Scott selected the five most popular algorithms: the similarity index, PBM, Euclidean distance, absolute value

distance, and the dot-product approach. They started by optimizing the intensity scaling and mass weighting factors of all the algorithms except PBM, for which they used the previously optimized values.[16] This optimization process is important since it de-emphasizes the value of the most abundant peaks, which are not always the most characteristic, and gives more importance to larger *m/z* values, which generally have greater significance in spectral identification.[1,25] These weight factors are dependent on the mass spectral library used.[26] Next, they determined the accuracy of the optimized algorithms as percentage correct as a function of rank in the hit list.

Using the criteria of first-hit recognition, the dot-product algorithm performed best with 75% correct, followed by the Euclidean distance at 72% correct, absolute value distance at 68%, PBM at 65%, and finally similarity index at 64%.[16] When expanding the criteria to include second- and third-hit identifications, all estimations increased by approximately 18%.[16] One of the main problems with these retrieval algorithms is that they entirely rely on the real compound's inclusion in the reference library, which cannot always be guaranteed. Additionally, even if the real compound is in the reference library, it must then be assigned a high match factor.

To account for both requirements, Stein developed a match factor based on the probability of a correct identification. Stein derived two spectral match factors from a weighted average of two comparison functions.[15] The first is a measure of the angle between the two spectra (Eqn. 3), similar to the dot-product approach.

$$F_1 = \frac{\sum M(A_L A_U)^{1/2}}{[\sum MA_L \sum MA_L]^{1/2}} \qquad \text{Eqn. 3}$$

Where L and U denote peaks in the library and unknown spectra, respectively, and for each peak, M is the mass-to-charge ratio and A is the abundance normalized to the base peak.

The second factor is based on the relative intensities of pairs of adjacent peaks present in the library and unknown spectra,[15]

$$F_2 = \left(\frac{1}{N_{U\&L}}\right) \sum_{i=2}^{N_{U\&L}} \left(\frac{A_{L,i}}{A_{L,i-1}}\right)^n \left(\frac{A_{U,i}}{A_{U,i-1}}\right)^{-n} \quad \text{Eqn. 4}$$

Where $N_{U\&L}$ is the number of peaks shared between the library and unknown spectra, and n = 1 if the first abundance ratio is less than the second, or n = – 1 if the opposite is true.

The match factor, MF, is obtained when both factors are combined.[15]

$$MF = \frac{1000}{N_U + N_{U\&L}} (N_U F_1 + N_{U\&L} F_2) \quad \text{Eqn. 5}$$

The scale for match factor ranges from 0, for no peaks in common, to 1000 for a perfect match.[15]

Over time, the match factor was adapted to obtain the Simple Similarity Search (SSS), which is defined in Eqn. 6 as:[17]

$$SSS\ (\boldsymbol{Q}, \boldsymbol{L}) = C \frac{\left(\sum_i \sqrt{Q_i} \times \sqrt{L_i}\right)^2}{\sum_i Q_i \times \sum_i L_i} \quad \text{Eqn. 6}$$

where Q and L are the vectorial representations of the queried and library spectra, respectively, Q_i and L_i are their corresponding abundances at unit mass i, and C is a constant that, for historical reasons, has a value of 999. Similar to MF, the range for SSS is from 0 to 999, where a score above 800 is considered "good", and a score below 700 is questionable.[17,27]

Today, the most widely used algorithm is probably the SSS, as it is implemented by the NIST Search Program.[27] However, this method still has its limitations when it comes to providing a reasonable match factor when the reference library does not contain the correct compound. To circumvent this drawback, Moorthy et al. developed a hybrid match factor that does not require the spectrum of a query sample to be included in the library.[17]

The Hybrid Similarity Search (HSS) is a natural extension of the SSS system that can match query peaks with a shifted library peak, in addition to a direct *m/z* match.[17] This shift is referred to

as *DeltaMass* (Δ_m) and it is the nominal mass difference between the query and library compound.[15] The hybrid match factor is given by:

$$HSS(\boldsymbol{Q}, \boldsymbol{L}, \Delta_m) = SSS(\boldsymbol{Q}, \boldsymbol{H}) \quad \text{Eqn. 7}$$

where H is the vector that contains the matching peak intensity information from the library and shifted peaks.

The main limitation of the HSS is the need for the molecular mass of the query compound, which is not always known. Acknowledging this drawback, Moorthy et al. provided a method to estimate the nominal mass of an unknown molecule.[17]

Koo et al. developed an approach based on the Discrete Fourier Transform (DFT) and Discrete Wavelet Transform (DWT), creating a composite method that was able to recover 4% more information than the dot-product algorithm optimized by Stein and Scott. Overall, the DFT/DWT-based composite approach correctly identified ~3% more spectra than Stein and Scott's algorithm.[16,25]

Different multivariate statistics have also been explored to discriminate spectrally similar compounds.[28,29] Bonetti focused on performing principal component analysis (PCA) followed by linear discriminant analysis (LDA) for the differentiation of positional isomers, achieving zero misclassifications for two different sets of isomers.[28] On the other hand, Setser et al. aimed their research at classifying emerging synthetic drugs according to their structural class. They used prior knowledge of the molecule's structure and PCA to select the variables that were then used to build two LDA models, which resulted in 93% and 86% successful classification rate, respectively.[29] Neither study included spectra collected on different instruments, which is the most challenging aspect of spectral identifications.

There has also been research into the development of a mass spectral equivalent of an error rate. Bodnar Willard et al. designed a way to calculate a random match probability (RMP) that the specific mass spectral fragmentation pattern occurred by chance. This estimation was done to spectra previously determined to be significantly similar, obtaining RMP values less than 10^{-29}.[30] However, these RMPs require that relative or normalized ion abundances are independently variable, which is not the case.[31,32]

Given the high vacuum conditions of electron ionization (EI) sources, EI spectra are the result of unimolecular dissociations.[33] The abundance of the fragments depends on four main factors:[33–35] 1) the internal energy distribution of the molecule prior to ionization, 2) the excitation energy accompanying the ionization event, 3) the observation time specific to the instrument, and 4) mass bias and spectral distortion caused by ion optics and instrument operation.[21,36] The fragmentation rates of a molecule can be modeled by the Rice–Ramsperger–Kassel–Marcus (RRKM) theory or quasi-equilibrium theory, which have very similar principles but different mathematical details.[37–41] RRKM theory describes the fragmentation pathways in terms of the enthalpy and entropy of activation associated with the transition states.[33,42] Fragmentation pathways are usually divided into two categories: rearrangements and direct bond cleavages. Rearrangements require a 'tight' transition state, which means they require a specific arrangement and geometry, usually close to those of the product. Tight transition states tend to have the lowest activation barriers and are typically favored at low internal energies.[33] In contrast, direct bond cleavages tend not to have specific conformational requirements. Instead, they have 'loose' transition states with higher activation barriers and therefore tend to be favored at high internal energies.[33] Therefore, fragments deriving from low energy, slow rearrangements are relatively more abundant at the lowest excitation energies and fragments from high-energy, fast cleavages

are relatively more abundant at higher internal energies. Each fragment ion abundance can be calculated at different internal energies at a fixed reaction time. Although branching ratios are non-linearly related to variations in internal energy or apparent observation time, branching ratios do display strong empirical linear correlations in replicate measurements.[31,32,43] These empirical linear correlations imply that certain ion ratios are relatively constant over a wide range of internal energies, and taking a weighted average of several ratios within a spectrum can be an effective way to extrapolate data across different internal energies, and thus, instruments. The relationship between RRKM/QET theories and the empirical correlations of a ~128 replicate spectra of cocaine from crime laboratories are described in detail in our recent publications.[32,43]

3. Materials and methods

3.1 Materials

Cocaine and ACS-grade methanol were purchased from Sigma Aldrich (St. Louis, MO). A stock solution of cocaine in methanol was prepared to obtain a final concentration of 200 ppm. Additionally, the NIST-EPA-NIH database, containing different compounds from different laboratories and instruments was used. Most of the replicate spectra of cocaine and its diastereomers from the NIST archive came from a select number of laboratories, including the NYC Police Laboratory, the NIST MS Data Center, the Defense and Civil Institute for Environmental Medicine, Canada, the Georgia Bureau of Investigation, and the Virginia Department of Forensic Science (see **Table 1**).

Table 1. Existing database of in-house and NIST spectra.

Name	Molecular formula	Molecular weight (g/mol)	Number of spectra
Ecgonine methyl ester	$C_{10}H_{17}NO_3$	199.25	69
Fentanyl	$C_{22}H_{28}N_2O$	336.5	216
Heroin	$C_{21}H_{23}NO_5$	369.4	158
Hydromorphone	$C_{17}H_{19}NO_3$	285.34	134
Methamphetamine	$C_{10}H_{15}N$	149.23	133
Pseudococaine	$C_{17}H_{21}NO_4$	303.35	8
Allococaine	$C_{17}H_{21}NO_4$	303.35	1
Pseudoallococaine	$C_{17}H_{21}NO_4$	303.35	2
Cocaine	$C_{17}H_{21}NO_4$	303.35	895

3.2 Gas chromatography-Mass spectrometry

3.2.1 Instrument parameters and data extraction

All samples were run on an Agilent Technologies (Santa Clara, CA, USA) 7890B GC system fitted with an HP-5MS column (30 m × 0.25 mm × 0.5 µm) and an Agilent 5977A mass spectrometer.

Table 2. GC-MS parameters for the analysis of cocaine.

Injection volume	1.0 µl
Split ratio	20:1
Carrier gas	Helium
Scan region (*m/z*)	30 - 350
Initial temperature	190 ºC
Initial hold time	0 min
Ramp rate	15 ºC/min
Final temperature	265 ºC
Final hold time	0 min
Solvent delay	1.5 min
Total separation time	5 min

The cocaine samples were analyzed using 27 different treatments stemming from the combination of three factors—repeller voltage, ion focus voltage, and electron ionization energy—each with three levels, as defined in **Table 3**.

Table 3. Different treatments for the GC-MS analysis of cocaine.

Treatment	Repeller voltage	Ion focus voltage	Electron ionization energy
1	20	70	67
2			70
3			72
4		90	67
5			70
6			72
7		110	67
8			70
9			72
10	30	70	67
11			70
12			72
13		90	67
14			70
15			72
16		110	67
17			70
18			72
19	40	70	67
20			70
21			72
22		90	67
23			70
24			72
25		110	67
26			70
27			72

Each treatment was run in triplicate. To accomplish randomized data acquisition, all samples were divided into three blocks, where each block contained all possible combinations exactly once (**Table 3**). Each block was independently randomized using Excel to ensure no bias when running the samples. The different parameters were set and saved to individual tune files that were then assigned their unique MS method. A methanol blank was run before each sample

to limit the possibility of carryover from previous samples. This design had 81 injections of cocaine resulting in an average of 7 scans per eluting peak for a total of 583 extracted spectra. Each scan was extracted using the 'Export 3D data' tool on ChemStation.

3.3 Data analysis

The MANOVA and consequent analysis of results were performed with Statistical Package for the Social Sciences (IBM SPSS, version 28). EASI was developed using a combination of Microsoft Excel for the data filtering and SPSS for building and testing the algorithms. SPSS was also used for calculations of metrics to determine the effectiveness of the algorithm, like mean absolute residuals, Pearson product-moment correlation (PPMC) coefficients, NIST scores, Mahalanobis distances and Euclidean distances.

3.3.1 Data selection and filtering

The extracted data from ChemStation was compiled into a master Excel spreadsheet that contained a unique ID number—comprised of a sequence of numbers associated with the treatment used, the block the sample belongs to and the scan number—the parameter conditions, the sequence order and the scan number. Next, the spectra underwent a data selection process that determined the 20 most abundant non-background ions in the cocaine database. Then, the data was filtered to remove poor quality spectra, which were defined as spectra with contamination (e.g., *m/z* 44, 73, 149, 210), high noise levels, and low-abundance spectra. This filtering process included removing any spectra whose base peak was less than 8,000 counts and ensuring that all ion abundances for the 20 selected *m/z* values were greater than 2,000 counts. After the data was filtered, all spectra were normalized relative to their corresponding base peaks at 100%. The base peak was typically, but not always, *m/z* 82.

3.3.2 Analysis of variance

The three factors (repeller voltage, ion focus voltage and electron ionization energy) and three levels (low, medium and high) for each factor resulted in a 3^3 full factorial design. To determine the effects of each parameter on the ion abundances, MANOVA was carried out using the master Excel file with the filtered, normalized data. Each top 20 *m/z* value was iteratively considered to be the response or dependent variable, and the three factors were the independent variables. The F-statistics and p-values were used to assess the significance of each independent variable and their possible interactions. In addition, the eta squared values of the factors and their interactions were calculated to quantify their contribution to the observed variance.

3.3.3 General linear model calculation

The EASI algorithm requires the prediction of 20 ion abundances using 20 linear models. To build one linear regression model, SPSS assigns one of the 20 most abundant peaks as the dependent, or response, variable and identifies the coefficients (β_n) of a subset of the remaining 19 independent variables to explain as much of the variance in the dependent variable as possible without overfitting, as described below. SPSS then reports estimates and ranges for each β_n value

$$\hat{x} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \ldots + \beta_n x_n \qquad \text{Eqn. 8}$$

where $\hat{x}$ is the predicted ion abundance, x_n represents each *m/z* value included in the model, and β_0 is the intercept. The predicted and measured ion abundances differ by a residual error, ε, which for multiple predictions should be randomly distributed if the model accurately explains the sources of variance in the data. The Normality of the residuals was demonstrated before and confirmed here.[32,43] In this project we evaluated two linear models, one that includes an intercept and one without (labeled EASI WO).

Each linear model is the result of the sequential addition or removal of independent variables based on the correlation between them and the response. The highly correlated covariates are added if they contribute significantly to the variance explained by the model ($p \leq 0.05$) and the least correlated covariates already included in the model are removed if they are no longer significant ($p \geq 0.10$). This process stops when all included covariates have a p-value less than 0.10 and all remaining covariates have a larger p-value than 0.05. This setup allows the final model to only include significant independent variables and reduce the risk of producing models that over-fit the data.

In addition to providing the β value for each linear model, SPSS also provides the predicted value of the dependent variable (based on the regression line) and the absolute residual to the measured value. To assess the normality of the residuals, we examined the frequency distribution of normalized residuals, scatter plots of normalized residuals versus normalized predicted values and probability-probability plots (P-P plots) of residuals.

3.3.4 Model predictions and assessment

The metrics used to assess the accuracy of the model are PPMC, mean absolute residuals (MAR), Euclidean distance values, NIST scores and Mahalanobis distances.

The PPMC values are defined by

$$r = \frac{n(\sum x\hat{x}) - (\sum x)(\sum \hat{x})}{\sqrt{\left[[n\sum x^2 - (\sum x)^2][n\sum \hat{x}^2 - (\sum \hat{x})^2]\right]}} \qquad \text{Eqn. 9}$$

Where r is the PPMC value, n is the sample size, $\hat{x}$ is the predicted ion abundance, and x is the measured ion abundance. The PPMC values are a measure of the linear correlation between two data sets, in this case, the predicted values and the measured values. These values range from

-1 to 1, where -1 is a perfect negative correlation and 1 is a perfect positive correlation. Although improbable, it is hypothetically possible to obtain a PPMC of 1 or -1 and still have constant or proportional differences between predicted and measured ion abundances. In this respect, forcing the linear regression line through the origin or using a simple dot product would be superior.

The next metric is the mean absolute residuals. The residuals are a measure of the pairwise error between the predicted and measured data. Thus, MAR is defined by:

$$MAR = \frac{\sum|\hat{x}_i - x_i|}{n} \qquad \text{Eqn. 10}$$

Where $\hat{x}_i$, x_i and n are as described above for PPMC. Unlike the PPMC values, the MAR uses the same units as the data being analyzed and ranges from zero to a theoretical maximum of 100 assuming the unlikely scenario of 20 predictions with the maximum error of 100% each. Instead, MARs tend to be greater than 5% for known negatives and less than 5% for known positives relative to the mean of all known positives (the consensus exemplar spectrum).

The third measure of assessment of the model is the Euclidean distance, described by:

$$d_{Euclid} = \sqrt{\sum(\hat{x}_i - x_i)^2} \qquad \text{Eqn. 11}$$

where $\hat{x}_i$ and x_i are defined as above. The Euclidean distance finds the straight-line distance between the multivariate data. Euclidean distances tend to be greater than 62 for known negatives and less than 9 for known positives and range from zero to a hypothetical upper limit of ~447 for 20 predictions of 100% error.

To calculate the NIST scores, the first step is to adjust the peak intensities to obtain weighted variables w:

$$w = [peak\ abundance]^{0.6} \cdot [peak\ mass]^3 \qquad \text{Eqn. 12}$$

where the exponents 3 and 0.6 correspond to the optimized values reported by Stein.[16] This metric is designed to decrease the emphasis on the most abundant peaks, which tend to be less diagnostic for a particular substance, and to enhance the relative significance of higher mass peaks since they are the least common and the most diagnostic.[16]

The next step to obtain the NIST scores is to calculate the dot product between the *w* values for both the query and the reference spectra, and finally multiply that by 999 to obtain a weighted score between 0 and 999, where 999 represents a perfect match.

An additional measure of dissimilarity used when presented with correlated variables is the Mahalanobis distance. The Mahalanobis distance represents the distance between a datapoint in multivariate space and the distribution of datapoints of a given dataset, in this case the training set. Unlike the Euclidean distance, the Mahalanobis distance considers the correlation between the variables and provides the distance in multivariate space.

$$d_{Mahal} = \sqrt{(x_i - \bar{x})^T \cdot \boldsymbol{C}^{-1} \cdot (x_i - \bar{x})} \qquad \text{Eqn. 13}$$

In this equation, x_i is an object vector, $\bar{x}$ is the arithmetic mean vector, T is the transpose matrix, and $\boldsymbol{C}$ is the sample covariance matrix. The Mahalanobis distances are calculated directly to the training set using the normalized spectra, with no additional linear modeling. Mahalanobis distances are not commonly used in mass spectrometry because relative ion abundances are assumed to be independently variable.[30,44,45] The assumption of channel independence contrasts many correlation studies that demonstrate strong correlations between branching ratios in replicate spectra.[31,32,43,46]

Additionally, to compare the EASI approach to the NIST consensus approach, we computed the same assessment values against a consensus spectrum, which in this case will be the average spectrum of all the cocaine spectra in the database. We will also use scatter plots,

frequency distribution plots, and binary classifiers to compare both algorithms. Performance as a binary classifier will be assessed through receiver operating characteristic (ROC) curves.[47–49]

A receiver operating characteristic (ROC) curve is a graphical representation of the sensitivity (y axis) and 1–specificity (x axis) of a binary classifier. The sensitivity, or true positive rate (TPR), corresponds to the proportion of correct positive classifications, whereas 1–specificity, or false positive rate (FPR), corresponds to the proportion of incorrect positive classifications. A ROC curve allows the user to use metrics as binary classifiers, in this case a positive or negative identification. A plot showing FPR versus TPR helps visualize the trade-off between true positives and false positives at a chosen threshold. Using the similarity and dissimilarity measures described above as continuous variables, the number of true positives (TPs), true negatives (TNs), false negatives (FNs) and false positives (FPs) were calculated at different thresholds using every mass spectra in the database sorted into training set and testing for the GLM-based algorithms and the consensus approach.

$$FPR = \frac{FP}{FP + TN} \quad \text{Eqn. 14}$$

$$TPR = Sensitivity = \frac{TP}{TP + FN} \quad \text{Eqn. 15}$$

$$TNR = Specificity = \frac{TN}{TN + FP} \quad \text{Eqn. 16}$$

$$FNR = \frac{FN}{FN + TP} \quad \text{Eqn. 17}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad \text{Eqn. 18}$$

In addition to the ROC curves, we also calculated the area under the curve (AUC) as this has been continuously used to assess the discriminatory power of a test. The area under the ROC curve (AUC) ranges from 0.5-1. A value of 0.5 indicates no better than guessing and a value of 1

indicates perfect classification. An AUC of less than 0.5 would indicate that the classification rule needs to be inverted to beat random odds.

4. Results and discussion

4.1 MANOVA analysis

The 27 sets of conditions tested in the full-factorial design of experiments resulted in an average of 22 spectra per condition. Summary statistics are provided in **Table 4** for *m/z* 94. Similar tables were produced for the 20 most abundant fragments of cocaine. The statistics include the mean, standard deviation, and number of observations (N) for the 27 combinations. **Table 4** shows a slight decrease in the average ion abundances when the repeller voltage increases from 20-30 V, but no discernable differences between repeller voltages of 30-40 V. The same trend is observed for *m/z* values in the middle range (77, 81, 83, 94, 96, 97 and 105). The *m/z* values in the low range (41, 42, 51, 55, 67 and 68) showed decreased ion abundances when going from low (20 V) to high (40V) repeller voltage (see **Table 29** in the appendix) and the ion abundances of the high *m/z* values (122, 152, 182, 183, 198, 272 and 303) increased with increasing repeller voltages (see **Table 30** in the appendix).

The ion abundances for *m/z* values in the low and middle range generally increased when the ion focus voltage increased from low to high. However, the abundances seemed to slightly decrease when the repeller voltage was set to high. These differences already show the impact of interactions between factors on the ion abundances. On the other hand, the ion abundances of the high *m/z* values increased and then decreased across the low, medium and high levels of the ion focus voltage.

Lastly, the ion abundances show a general decrease when increasing the EI energy. This behavior is mostly observed in *m/z* values in the low and middle ranges. For the high *m/z* values, there is no significant change.

In general, the within-group variance was not significantly smaller than the between-group variance, resulting in conditions that are not significantly different. This behavior demonstrates the small variance observed throughout the experiment.

These overall descriptive statistics highlight the differences the individual parameters can have depending on the *m/z* values, in addition to indicating the interaction between effects might be significant. However, the small variance and overlap between configurations can obscure these trends or make them irrelevant.

Table 4. Descriptive statistics for all 27 possible combinations of factors for *m/z* 94.

m/z	Repeller voltage	Ion focus voltage	EI energy	Average	Std. Deviation	N
94	20	70	65	37.268	1.3894	12
			70	36.538	1.4201	14
			80	33.593	1.6134	18
			Total	35.532	2.2084	44
		90	65	43.739	0.9091	12
			70	40.772	2.5659	15
			80	38.194	2.0572	17
			Total	40.585	2.9964	44
		110	65	45.949	2.0913	11
			70	44.493	2.4741	12
			80	41.038	2.4923	16
			Total	43.486	3.1580	39
	30	70	65	34.695	2.4995	23
			70	34.035	2.0162	22
			80	34.128	1.6212	22
			Total	34.292	2.0736	67
		90	65	37.379	2.5559	26
			70	36.786	2.3901	25
			80	35.695	2.5642	30
			Total	36.572	2.5794	81
		110	65	41.425	2.2640	23
			70	40.951	2.4670	23
			80	39.164	2.0590	25
			Total	40.475	2.4418	71
	40	70	65	38.933	2.5916	22
			70	38.926	1.7321	20
			80	38.080	1.4666	19
			Total	38.665	2.0257	61
		90	65	38.476	2.2905	29
			70	37.479	3.0405	31
			80	37.349	2.4552	29
			Total	37.762	2.6440	89
		110	65	41.271	2.4879	28
			70	40.004	3.1062	29
			80	38.395	2.6640	30
			Total	39.857	2.9802	87

The main advantage of performing MANOVA over multiple ANOVAs is the ability to measure the effects of independent factors on multiple dependent variables. Consequently, a MANOVA can assess the variance contribution of the interactions between factors. Additionally, MANOVA can decrease the risk of type I error, which is the rejection of the null hypothesis when it is true.

Table 5 shows a summary of the MANOVA results. The reported F-statistics and p-values are based on the Wilks' Lambda statistical test. Significant factors are characterized by larger F-statistics that provide p-values less than 0.05 (for 95% confidence). By these standards, all the parameters were deemed significant except for the interaction between ion focus and EI energy. **Table 5** also shows the eta squared values, which represent the percent variance explained by that source. Most of the variance is explained by the intercept, which is unrelated to any of the studied factors. Of the three factors studied, the repeller voltage had the largest effect, but only explained ~3% of the observed variance. The lowest eta squared value corresponds to the ion focus and EI energy interaction. This double interaction accounts for less than 0.001% of the variance observed, which is consistent with the higher p-value obtained.

Table 5. Summary of MANOVA results showing the F-statistic, p-value and eta squared for each source.

Source	**F-statistic**	**p-value**	**Eta squared**
Intercept	56439.560	0	93.731
Repeller	114.906	0	2.792
Ion Focus	67.957	<0.001	0.404
EI Energy	23.732	<0.001	0.057
Repeller * Ion Focus	14.139	<0.001	0.131
Repeller * EI Energy	9.412	<0.001	0.086
Ion Focus * EI Energy	1.031	0.406	0.007
Repeller * Ion Focus * EI Energy	1.893	<0.001	0.014

Table 6 shows that the behaviors outlined previously can be broken down further if we consider the *m/z* values as a scale with low, middle, and high values as before. Now, the eta squared values allow us to analyze mass bias for each parameter, including the interactions. Again, most of the variance is not explained by the studied factors. Although some of the factors and interactions are significant, the eta squared values are still small relative to the overall variance. The repeller voltage is most impactful for high *m/z* values and explains about 7% of the observed variance in the relative ion abundances. The intercept, which explains most of the variance, therefore explains less variance for high *m/z* values.

Table 6. Averaged eta squared values by source and m/z values.

	Average eta squared values			
Source	**Total**	**Low *m/z* (*m/z* 44-68)**	**Middle *m/z* (*m/z* 77-107)**	**High *m/z* (*m/z* 122-303)**
Intercept	93.73	97.33	98.87	85.44
Repeller	2.792	0.555	0.147	7.372
Ion Focus	0.404	0.589	0.205	0.474
EI Energy	0.057	0.050	0.045	0.082
Repeller * Ion Focus	0.131	0.164	0.056	0.186
Repeller * EI Energy	0.086	0.189	0.036	0.051
Ion Focus * EI Energy	0.007	0.006	0.004	0.011
Repeller * Ion Focus * EI Energy	0.014	0.028	0.005	0.011

The effect of the repeller mass bias is more evident in a bivariate plot of percent change in relative ion abundance per volt when going from a repeller voltage of 20 V to 40 V versus *m/z* value. **Figure 1** shows a linear increase of the percent change in ion abundance as the *m/z* values grow larger. For *m/z* values between 44 and 68, an increase in repeller voltage decreases the relative abundance. For *m/z* values between 77-105, the repeller voltage has very little effect on the relative ion abundance. For m/z values greater than *m/z* 122, the increase in repeller voltage translates into an increase in relative abundance. The line of best fit in this case can explain 95% of the behavior displayed by the data.

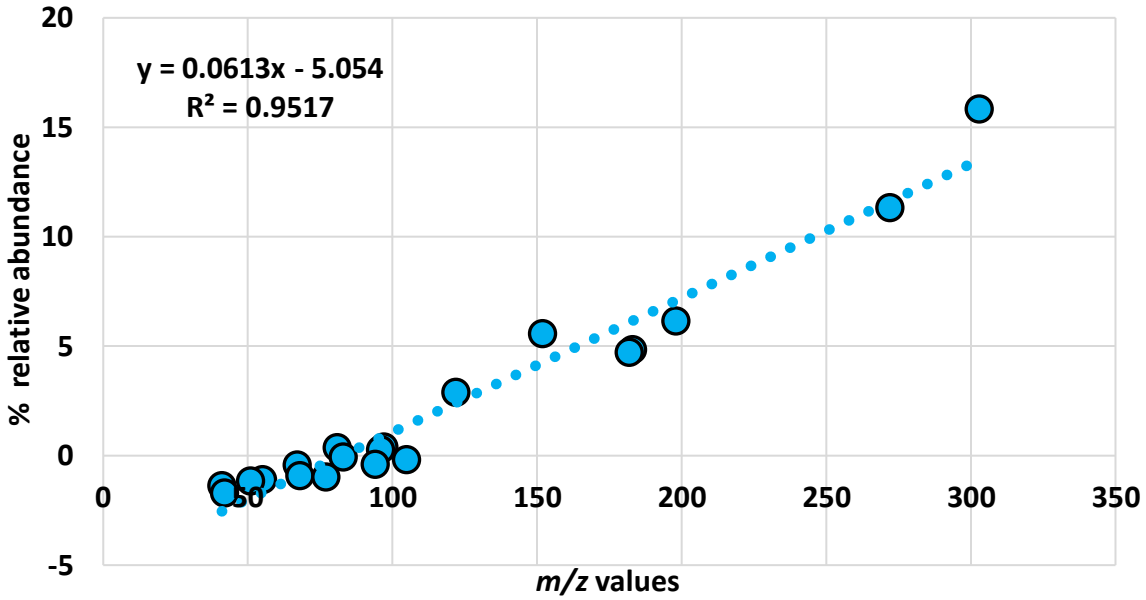

Figure 1. Percent change in relative ion abundance as a function of *m/z* when going from a low (20 V) to a high (40 V) repeller voltage.

The repeller causes mass bias in this experiment because it was altered independently of the focus lens. When the instrument autotunes to provide unbiased mass spectra (of the calibration gas, perfluorotributylamine, PFTBA), the repeller and focus lens are both optimized to simultaneously enhance the extraction efficiency while negating the incurred mass bias.[43]

Previous studies by the Jackson group also explored the effects of column flow rate, ion source temperature and transfer line temperature on the top 10 relative ion abundances of various drugs using MANOVA. Column flow rate and ion source temperature were shown to be significant, with eta squared values of 0.15 and 0.73 for cocaine.[50] This same study monitored the vacuum chamber pressure while analyzing perfluorotributylamine (PFTBA) to determine the effects of natural pressure fluctuations on relative ion abundances. The results showed that the pressure explained up to 84% of the variance in the relative abundance of *m/z* 219 (**Figure 2**).[50]

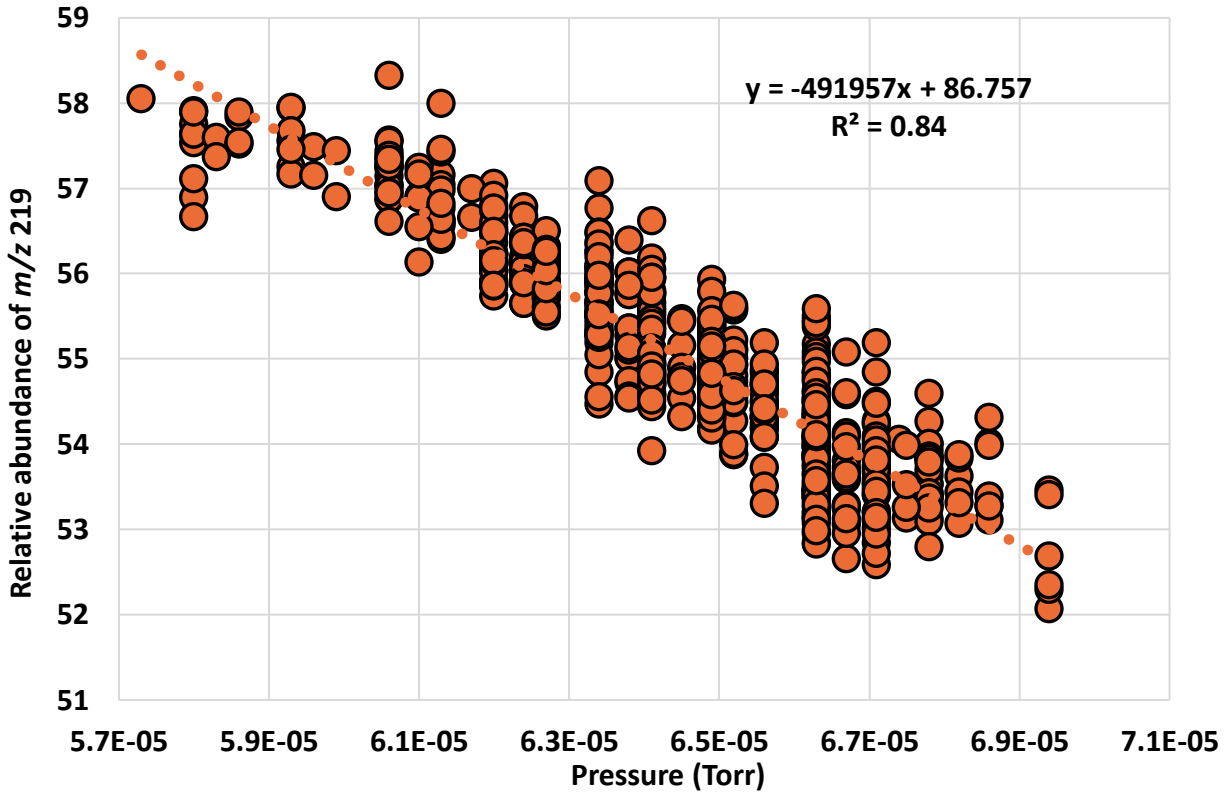

Figure 2. Relative ion abundance of PFTBA versus high vacuum (Torr) for *m/z* 219. The pressure randomly drifted over a ~2-min period during these observations.

To explore if this trend is reproducible with cocaine, the high vacuum was tracked while cocaine was continuously analyzed in manual tune mode. This setup only allowed the visualization and recording of three *m/z* values at a time. We therefore selected abundant ions that spanned the mass range and contained the important molecular ion; *m/z* 82, *m/z* 182 and *m/z* 303. Consequently, the high vacuum was monitored for 2 minutes and 20 seconds while the ion abundances were recorded every 5 seconds. Each ion abundance was then normalized to the most abundant peak of the three, in this case m/z 82. These results are summarized in **Table 7**.

Table 7. Pressure, absolute abundances, and relative abundances of cocaine ions *m/z* 82, 182 and 303 collected every 5 seconds for 2 minutes and 20 seconds. The pressure drifted without control during this period.

Time (sec)	Absolute abundance *m/z* 82	Absolute abundance *m/z* 182	Absolute abundance *m/z* 303	Relative abundance *m/z* 82	Relative abundance *m/z* 182	Relative abundance *m/z* 303	Pressure (Torr)
0	54263	29334	4768	100	54.1	8.8	5.42E-05
5	53696	28028	4498	100	52.2	8.4	5.42E-05
10	61282	32671	5532	100	53.3	9.0	5.42E-05
15	62814	34462	5850	100	54.9	9.3	5.42E-05
20	70891	39467	6280	100	55.7	8.9	5.42E-05
25	80589	44046	7554	100	54.7	9.4	5.42E-05
30	86912	47878	7942	100	55.1	9.1	5.42E-05
35	94294	54086	9091	100	57.4	9.6	5.42E-05
40	106418	59576	9830	100	56.0	9.2	5.42E-05
45	110607	63602	10069	100	57.5	9.1	5.45E-05
50	121134	67150	11382	100	55.4	9.4	5.48E-05
55	117389	65432	11083	100	55.7	9.4	5.54E-05
60	110942	61771	10423	100	55.7	9.4	5.54E-05
65	114905	62965	10277	100	54.8	8.9	5.54E-05
70	113411	61954	10132	100	54.6	8.9	5.60E-05
75	115911	61991	10182	100	53.5	8.8	5.70E-05
80	117660	63130	10611	100	53.7	9.0	5.80E-05
85	120493	64075	10372	100	53.2	8.6	5.86E-05
90	117057	61019	10034	100	52.1	8.6	5.96E-05
95	126555	65606	10774	100	51.8	8.5	6.06E-05
100	128886	67333	11183	100	52.2	8.7	6.13E-05
105	135992	68333	11308	100	50.2	8.3	6.24E-05
110	139555	69243	11672	100	49.6	8.4	6.34E-05
115	141302	68761	11302	100	48.7	8.0	6.49E-05
120	145778	69951	11226	100	48.0	7.7	6.71E-05
125	146991	69218	11359	100	47.1	7.7	6.86E-05
130	148611	66460	11119	100	44.7	7.5	7.10E-05
135	143595	62974	10200	100	43.9	7.1	7.42E-05
140	143545	60270	9894	100	42.0	6.9	7.76E-05

A scatter plot of relative abundance of *m/z* 182 is the dependent variable versus pressure as the independent variable resulted in a linear plot with a negative slope with a coefficient of

determination of 0.93 (**Figure 3**). This means that the high vacuum explains 93% of the variance observed in the relative abundance of the peak in question, which is in close agreement with the percent variance that was not explained by the controlled factors.

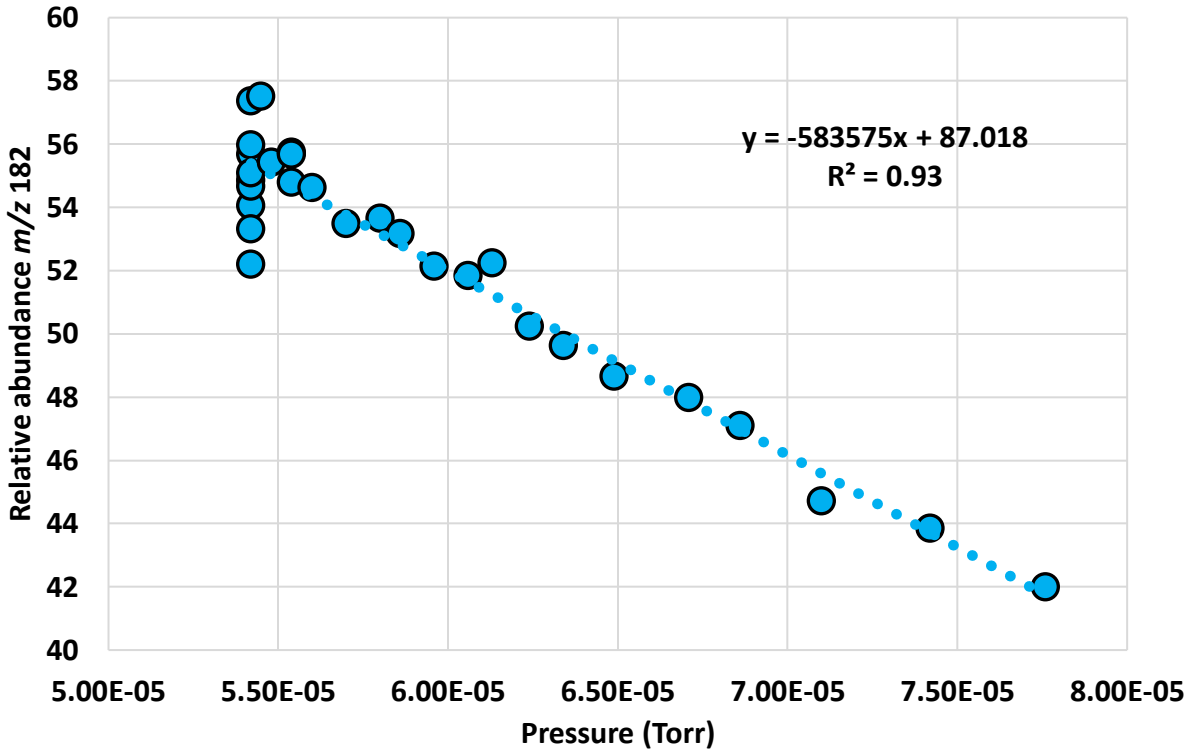

Figure 3. Relative abundance of cocaine peak *m/z* 182 versus the uncontrolled high vacuum pressure (Torr).

Similarly, a linear regression using the relative abundance of *m/z* 303 also revealed a negative correlation between the high vacuum and the 303 peak (**Figure 4**). The determination coefficient was 0.878, which represents 87.8% of the variation of the relative abundance of *m/z* 303 that can be attributed to the high vacuum.

These results suggest that the fluctuations in the high vacuum is responsible for most of the unexplained variance in the major peaks of cocaine.

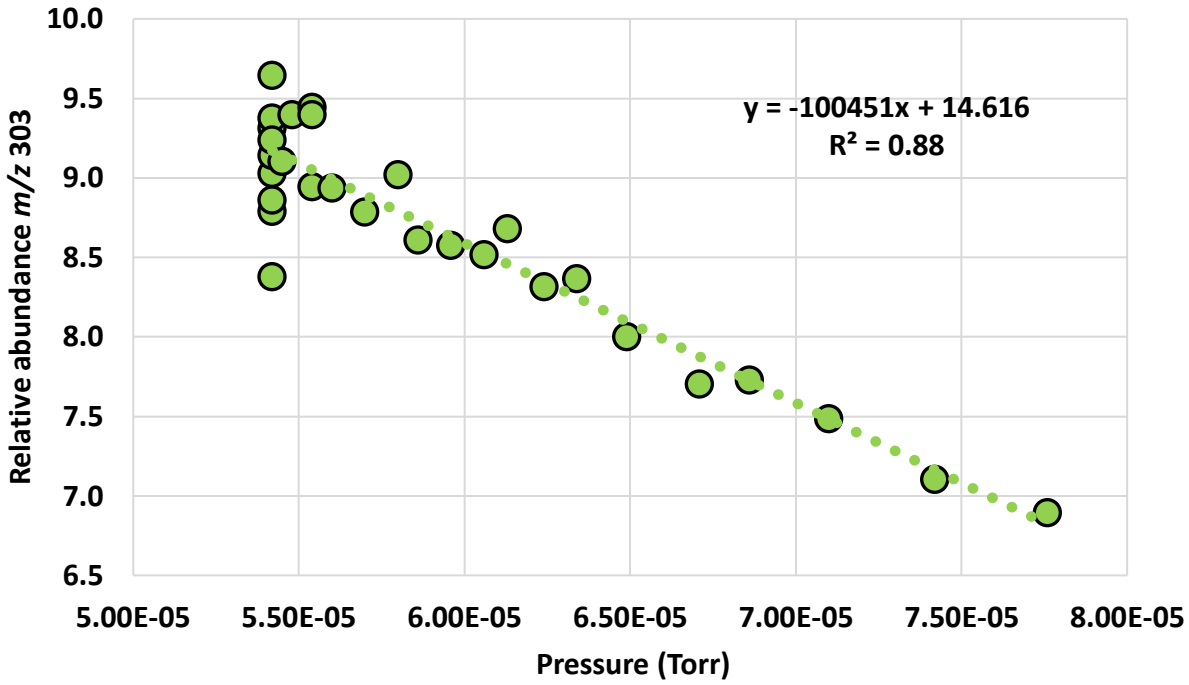

Figure 4. High vacuum (Torr) vs. relative abundance of cocaine peak *m/z* 303.

Additionally, **Figure 5** and **Figure 6** show that there is no significant correlation between the absolute and relative abundances of each of the analyzed ions. This means that the absolute abundances are not increasing and decreasing together, so the effect is not caused by self-chemical-ionization in the source.

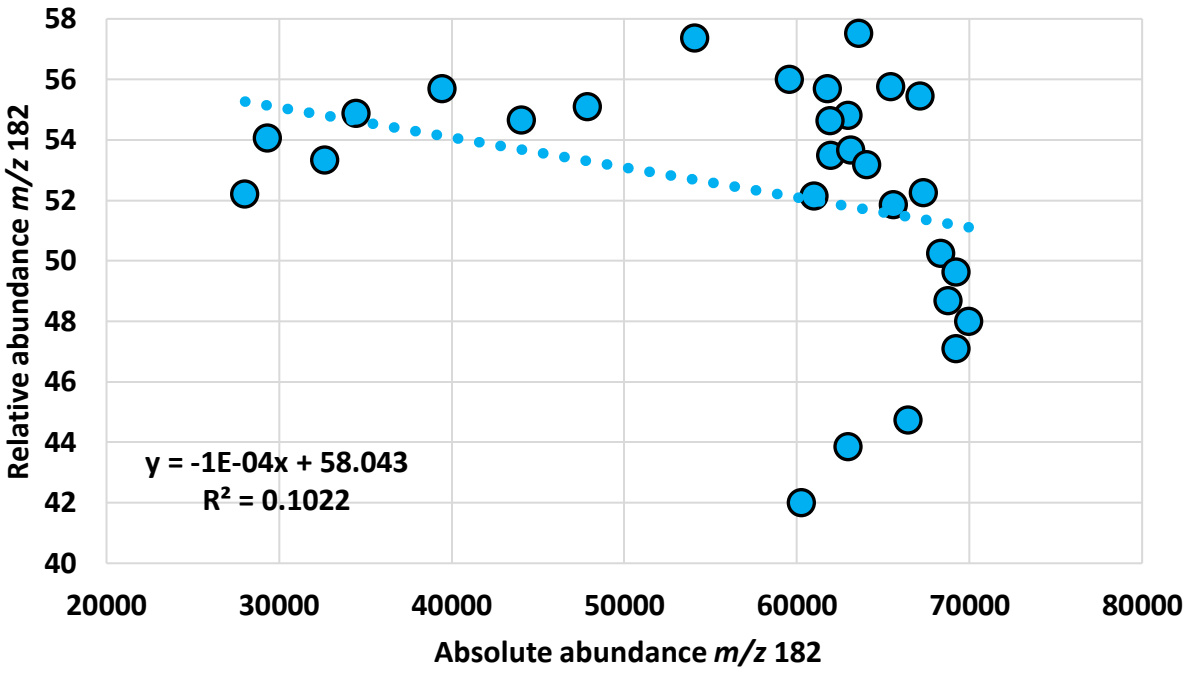

Figure 5. Absolute abundance vs. relative abundance of *m/z* 182.

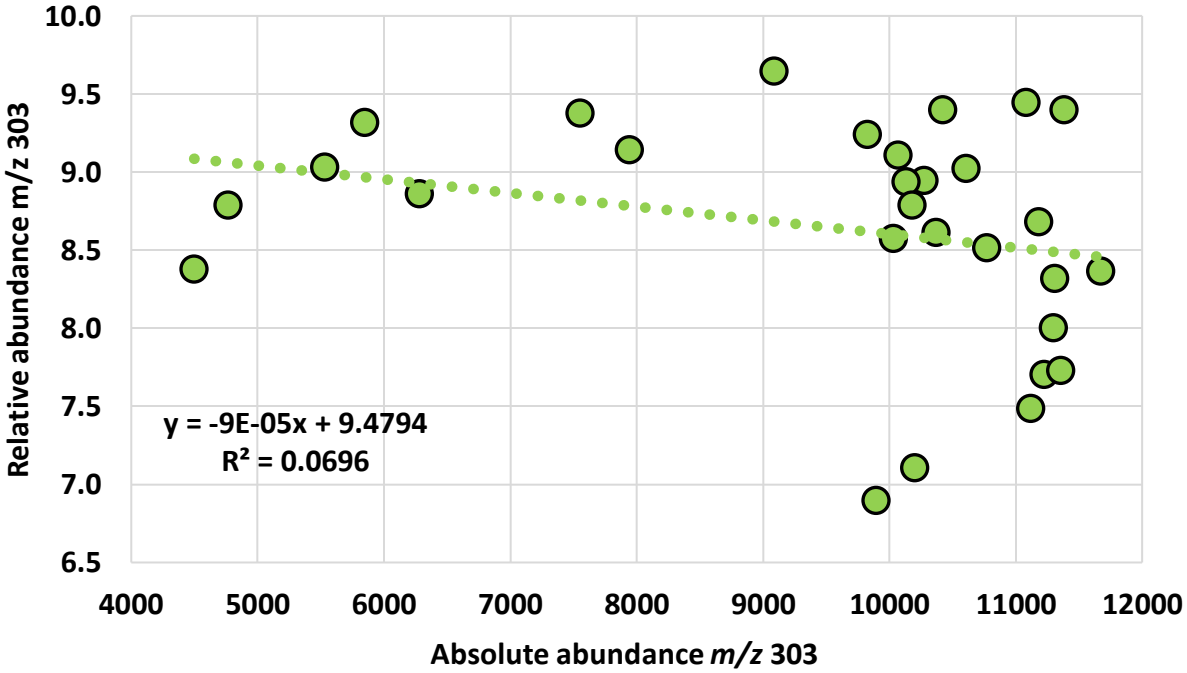

Figure 6. Absolute abundance vs. relative abundance of *m/z* 303.

Figure 7 shows the exponential increase in pressure as time passes. However, this behavior is not replicated by the absolute abundances, which suggests the increase in pressure is not caused by cocaine, but by random drift in the base pressure of the instrument. These observations are related to classical experiments in the early development of mass spectrometry in which precursor or intermediate ions were found to fragment after being accelerated out of the ion source. The post-extraction fragmentation either occurs naturally—because of the long lifetime of the activated ion—or because of collisions with residual gasses during flight.[34] Such long-lived intermediates would often show up as metastable peaks/ions in magnetic sector instruments. The current observations support the hypothesis that random drift in the base pressure of the instrument could be inducing extra collisional activation as the ions migrate the ion optics en route to detection.

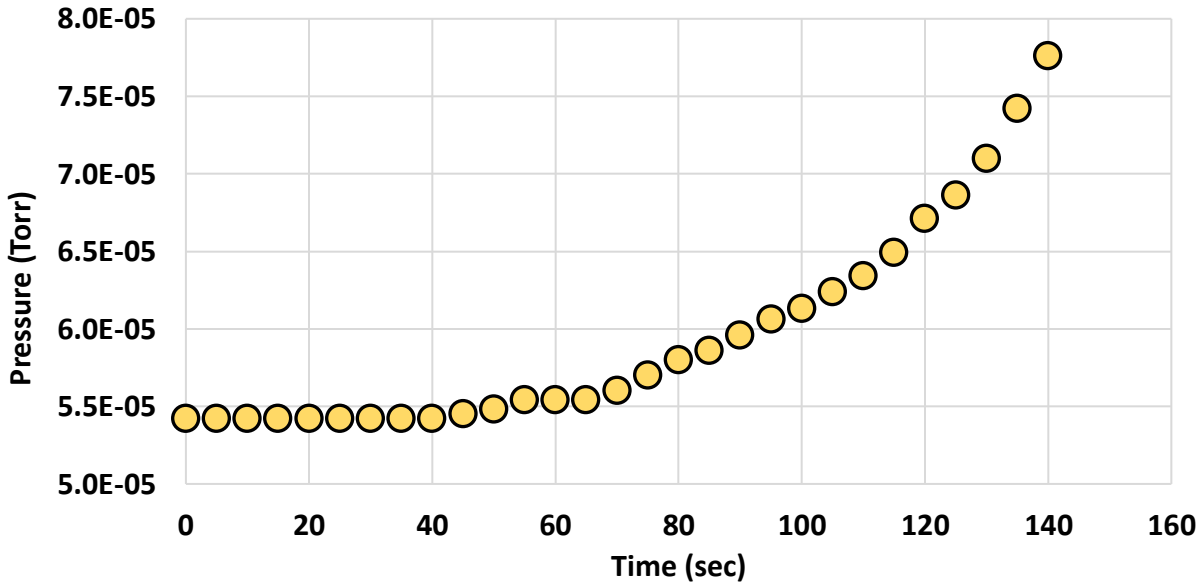

Figure 7. Pressure (Torr) vs. time while continuously analyzing cocaine.

4.2 GLM modeling

The data filtration of the collected spectra resulted in the following top 20 most abundant *m/z* values: 82, 182, 94, 77, 83, 105, 96, 42, 81, 51, 97, 303, 122, 198, 183, 55, 41, 68, 67, 272, and 152. These *m/z* values include the base peak for cocaine at 82 and the molecular ion at 303.

Although in this case the base peak was always *m/z* 82, cocaine can also have the base peak at *m/z* 182.[32,43] After the data selection and filtering process, the total number of reasonable quality cocaine spectra was 583. Of those, 389 made up the training set and 194 were used as a validation or test set.

The correlation tables (**Table 8** and **Table 9**) show the pairwise correlation coefficients for 19 of the ion abundances. The peak at *m/z* 82 was excluded because this value is always the base peak at 100%, so has no variance. Absolute correlation coefficients with a value greater than 0.6 or less than -0.6 indicate a significant correlation, which can be a direct relationship (positive correlation coefficients) or an inverse relationship (negative correlation coefficients). The correlation coefficients were significant for most of the pair-wise comparisons, which confirms the validity of the use of a multivariate linear regression approach. These findings are qualitatively similar to a completely different training set of 128 cocaine spectra from an operational crime laboratory.[32]

Table 8 shows highly correlated variables, like *m/z* 182 and *m/z* 183 (R=0.957), while also showing some highly inversely correlated variables, like *m/z* 83 and *m/z* 152 (R=0.782). These behaviors establish a trend where ion abundances close in *m/z* value tend to increase or decrease together (e.g., *m/z* 41 and *m/z* 42) and ion abundances with relatively different *m/z* values tend to show less correlation or inverse correlation. This correlation analysis presents a good indicator of which independent ions will be included in the models, i.e., the ions with the highest correlation coefficient have a higher probability of being included in the general linear models.

Table 8. Part one of the bivariate Pearson correlations between the top 20 most abundant ions of cocaine based on the training set (n=389).

	41	**42**	**51**	**55**	**67**	**68**	**77**	**81**	**83**	**94**
41	1.000	0.692	0.791	0.742	0.630	0.664	0.554	0.345	-0.381	0.169
42	0.692	1.000	0.772	0.838	0.140	0.695	0.861	-0.132	0.195	0.571
51	0.791	0.772	1.000	0.745	0.480	0.731	0.770	0.159	-0.198	0.464
55	0.742	0.838	0.745	1.000	0.333	0.693	0.724	0.058	-0.011	0.386
67	0.630	0.140	0.480	0.333	1.000	0.430	0.173	0.557	-0.770	-0.124
68	0.664	0.695	0.731	0.693	0.430	1.000	0.714	0.136	-0.190	0.446
77	0.554	0.861	0.770	0.724	0.173	0.714	1.000	-0.203	0.148	0.774
81	0.345	-0.132	0.159	0.058	0.557	0.136	-0.203	1.000	-0.607	-0.322
83	-0.381	0.195	-0.198	-0.011	-0.770	-0.190	0.148	-0.607	1.000	0.329
94	0.169	0.571	0.464	0.386	-0.124	0.446	0.774	-0.322	0.329	1.000
96	-0.372	0.085	-0.122	-0.144	-0.663	-0.174	0.165	-0.494	0.727	0.529
97	-0.484	-0.068	-0.299	-0.242	-0.681	-0.319	-0.032	-0.490	0.737	0.349
105	0.452	0.500	0.607	0.451	0.358	0.526	0.710	0.027	-0.168	0.728
122	0.135	-0.449	0.005	-0.237	0.627	-0.041	-0.334	0.594	-0.771	-0.287
152	0.114	-0.440	0.012	-0.225	0.623	-0.025	-0.303	0.570	-0.782	-0.230
182	-0.356	-0.539	-0.253	-0.530	-0.011	-0.316	-0.303	0.078	-0.189	0.130
183	-0.366	-0.519	-0.268	-0.526	-0.066	-0.319	-0.308	0.081	-0.132	0.128
198	-0.347	-0.566	-0.264	-0.524	0.040	-0.294	-0.337	0.150	-0.241	0.073
272	-0.212	-0.529	-0.190	-0.447	0.217	-0.210	-0.305	0.245	-0.423	-0.008
303	-0.215	-0.543	-0.196	-0.455	0.241	-0.222	-0.317	0.251	-0.447	-0.036

Table 9. Part two of the bivariate Pearson correlations between the top 20 most abundant ions of cocaine based on the training set (n=389).

	96	**97**	**105**	**122**	**152**	**182**	**183**	**198**	**272**	**303**
41	-0.372	-0.484	0.452	0.135	0.114	-0.356	-0.366	-0.347	-0.212	-0.215
42	0.085	-0.068	0.500	-0.449	-0.440	-0.539	-0.519	-0.566	-0.529	-0.543
51	-0.122	-0.299	0.607	0.005	0.012	-0.253	-0.268	-0.264	-0.190	-0.196
55	-0.144	-0.242	0.451	-0.237	-0.225	-0.530	-0.526	-0.524	-0.447	-0.455
67	-0.663	-0.681	0.358	0.627	0.623	-0.011	-0.066	0.040	0.217	0.241
68	-0.174	-0.319	0.526	-0.041	-0.025	-0.316	-0.319	-0.294	-0.210	-0.222
77	0.165	-0.032	0.710	-0.334	-0.303	-0.303	-0.308	-0.337	-0.305	-0.317
81	-0.494	-0.490	0.027	0.594	0.570	0.078	0.081	0.150	0.245	0.251
83	0.727	0.737	-0.168	-0.771	-0.782	-0.189	-0.132	-0.241	-0.423	-0.447
94	0.529	0.349	0.728	-0.287	-0.230	0.130	0.128	0.073	-0.008	-0.036
96	1.000	0.797	0.117	-0.442	-0.436	0.293	0.333	0.218	0.031	-0.013
97	0.797	1.000	-0.052	-0.420	-0.409	0.286	0.338	0.228	0.033	-0.003
105	0.117	-0.052	1.000	0.175	0.213	0.216	0.171	0.175	0.222	0.216
122	-0.442	-0.420	0.175	1.000	0.935	0.570	0.523	0.610	0.710	0.723
152	-0.436	-0.409	0.213	0.935	1.000	0.616	0.564	0.660	0.762	0.789
182	0.293	0.286	0.216	0.570	0.616	1.000	0.957	0.971	0.919	0.902
183	0.333	0.338	0.171	0.523	0.564	0.957	1.000	0.943	0.871	0.842
198	0.218	0.228	0.175	0.610	0.660	0.971	0.943	1.000	0.928	0.913
272	0.031	0.033	0.222	0.710	0.762	0.919	0.871	0.928	1.000	0.977
303	-0.013	-0.003	0.216	0.723	0.789	0.902	0.842	0.913	0.977	1.000

Table 10 summarizes the 19 general linear regression models when the model is allowed to include an intercept. There is no regression model for *m/z* 82 since it is always 100 and thus, can be described by a constant. For the same reason, *m/z* 82 was not selected as one of the covariates used for the models. **Table 10** shows the final selected covariates by the stepwise process using SPSS. The coefficients were stored in real-time and applied to all the selected mass spectra (e.g., training set spectra, validation spectra, known negative spectra).

Table 10. Model summary for the 19 regression models built including a constant. The summary includes the covariates that made it to the final model as part of the stepwise process.

Dependent *m/z*	**No. of stepwise additions and removals**	**Covariates included in the final model**	**R squared of final model**
42	5	51, 55, 77, 152, 303	0.894
51	5	42, 67, 77, 81, 122	0.793
55	8	42, 51, 68, 83, 152, 182	0.777
67	9	51, 55, 77, 83, 94, 96, 122, 152, 183	0.799
68	9	42, 55, 77, 81, 83	0.656
77	13	42, 51, 67, 68, 81, 94, 97, 105, 152, 182, 303	0.922
81	7	122, 152, 182, 183, 198	0.489
83	6	67, 96, 97, 122, 152, 272	0.850
94	11	51, 67, 68, 77, 96, 105, 122, 182, 198, 272, 303	0.876
96	9	51, 67, 83, 94, 97, 152, 182, 183, 303	0.815
97	4	51, 83, 96, 183	0.759
105	7	42, 67, 68, 77, 94, 122, 303	0.803
122	9	42, 67, 81, 94, 105, 152, 182, 183, 303	0.901
152	11	42, 55, 67, 81, 83, 96, 122, 183, 198, 272, 303	0.933
182	13	51, 55, 68, 77, 81, 94, 96, 105, 122, 183, 198, 272, 303	0.975
183	6	81, 97, 182, 198, 272, 303	0.930
198	6	42, 68, 81, 182, 183, 272	0.958
272	7	55, 83, 122, 152, 183, 198, 303	0.967
303	5	55, 81, 122, 152, 272	0.962

Similar to the summary for the regression models including an intercept, **Table 11** shows the summary for models without an intercept, including *m/z* 82 since there is no constant that can model it now.

Table 11. Model summary for the 20 regression models built excluding a constant. The summary includes the covariates that made it to the final model as part of the stepwise process.

Dependent *m/z*	No. of stepwise additions and removals	Covariates included in the final model	R squared of final model
42	11	51, 55, 68, 77, 82, 96, 105, 122, 198	0.997
51	5	42, 67, 77, 83, 122	0.995
55	15	42, 51, 68, 82, 83, 152, 182	0.997
67	12	51, 55, 77, 82, 83, 94, 96, 122, 152, 183	0.995
68	5	42, 55, 67, 77, 81	0.995
77	12	42, 51, 67, 68, 81, 94, 105, 152, 182, 303	0.999
81	8	82, 122, 152, 182, 183, 198	0.997
82	12	42, 51, 67, 68, 81, 83, 96, 105, 122, 183	0.999
83	9	67, 82, 96, 97, 122, 152, 272	0.999
94	10	51, 67, 68, 77, 96, 105, 122, 182, 198, 272	0.999
96	9	42, 67, 82, 83, 94, 97, 152, 182, 303	0.999
97	4	67, 83, 96, 183	0.997
105	7	55, 67, 68, 77, 94, 122, 303	0.999
122	8	42, 67, 81, 94, 105, 152, 182, 303	0.994
152	11	42, 55, 67, 81, 83, 96, 122, 183, 198, 272, 303	0.991
182	13	51, 55, 68, 77, 81, 94, 96, 105, 122, 183, 198, 272, 303	0.998
183	7	67, 97, 122, 182, 198, 272, 303	0.995
198	5	42, 68, 182, 183, 272	0.996
272	8	68, 83, 182, 183, 198, 303	0.993
303	8	55, 122, 152, 182, 183, 272	0.991

Comparing **Table 10** and **Table 11**, we can see that the models that include an intercept have the same number of variables or less, never more, than the models that exclude an intercept. This correlation analysis presents a good indicator of which covariate ions are likely to be included in the final models because ions with the highest correlation coefficient are more likely to be included in the model.

The R squared values shown in **Table 10** and **Table 11** demonstrate that the spectral variance is not random, as implicitly assumed by other algorithms.[30] On average, EASI can

describe at least 84.5% (regression with intercept) of the training set variance. These high R squared values illustrate how EASI can account for inter-instrument and inter-laboratories variations.[32,43]

Table 12 and **Table 13** show summaries of the resulting unstandardized coefficients generated through general linear modeling with and without an intercept, respectively. The coefficients in **Table 12** share significant qualitative similarities with the coefficients determined from a completely different training set in which the 128 cocaine replicates were collected over a 6-month period in an operational crime laboratory and the three factors in question were not deliberately manipulated as they were here. Despite the differences in instruments and controlled experimental perturbations, many similarities are present in the linear models. For example, in both models, the GLMs for *m/z* 303 include a small intercept of 2-3%, they have large coefficients of 2-3 for *m/z* 272, and they both include a negative coefficient for *m/z* 81 on the order of -0.2 to -0.5. The models do contain some differences, however, as one might expect for a dataset with such highly correlated variables.

Table 12. Summary of the unstandardized coefficients for 19 general linear regression models using the abundance of each *m/z* value as a dependent variable and the remaining 18 abundances as possible covariates. This model also includes a constant.

Dependent *m/z* value	Unstandardized coefficients for independent *m/z* values																			
	βo	42	51	55	67	68	77	81	83	94	96	97	105	122	152	182	183	198	272	303
42	-4.44		0.57	1.47			0.31								-0.41					-0.06
51	-5.10	0.25			0.27		0.15	0.18						0.20						
55	2.28	0.13	0.06			0.12			0.03						0.14	-0.02				
67	6.11		0.08	0.10			0.06		-0.06	-0.04	-0.08			0.15	0.15		-0.11			
68	2.60	0.04		0.13			0.07	0.08	-0.05											
77	4.68	0.21	0.54		0.30	0.50		-0.48		0.58		-0.25	0.30		-0.39	-0.09				0.21
81	10.87													0.19	0.16	-0.06	0.14	0.17		
83	25.64				-0.46						0.33	0.89		-0.16	-0.38				-0.18	
94	2.15		-0.22		-0.28	0.52	0.44				0.17		0.37	-0.19		0.10		0.37	-0.33	-0.12
96	7.09		0.12		-0.26				0.20	0.10		0.33			-0.24	0.07	0.16			-0.07
97	-0.99		-0.04						0.19		0.15						0.18			
105	1.79	0.10			0.35	-0.36	0.24			0.39				0.37						0.09
122	3.70	-0.05			0.33			0.20	-0.06	-0.11			0.09		0.64	0.05				-0.09
152	1.99	-0.04		0.22	0.19			0.11	-0.11		-0.09			0.40			0.13	0.15	-0.27	0.18
182	5.10		0.53	-0.84		-1.06	-0.39	-0.89		0.44	0.61		0.31	0.37			2.15	2.79	0.95	0.40
183	-1.84							0.08				0.14				0.08		0.25	0.16	-0.08
198	-0.88	-0.04				0.13		0.07								0.08	0.21		0.16	
272	0.77			0.08					-0.05					0.06	-0.12		0.17	0.16		0.32
303	2.94			-0.23				-0.21						-0.20	0.56				2.06	

Table 13. 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. This model does not include a constant.

Dependent *m/z* value	Unstandardized coefficients for independent *m/z* values																			
	42	51	55	67	68	77	81	82	83	94	96	97	105	122	152	182	183	198	272	303
42		0.57	1.31		0.38	0.21		-0.07			0.19		0.11	-0.30				-0.39		
51	0.27			0.22		0.13			-0.07					0.19						
55	0.13	0.06			0.12			0.02	0.03						0.13	-0.02				
67		0.08	0.11			0.06		0.06	-0.06	-0.04	-0.08			0.15	0.15		-0.12			
68	0.03		0.11	0.14		0.07	0.10													
77	0.19	0.53		0.38	0.57		-0.36			0.59			0.32		-0.37	-0.10				0.21
81								0.11						0.19	0.16	-0.06	0.14	0.17		
82	-0.20	-0.25		1.83	1.21		1.84		1.22		0.49		0.27	0.61			-0.43			
83				-0.47				0.26			0.33	0.88		-0.16	-0.38				-0.18	
94		-0.23		-0.23	0.58	0.44					0.26		0.36	-0.16		0.09		0.35	-0.57	
96	0.08			-0.27				0.07	0.21			0.33	0.07		-0.24	0.08	0.16			-0.09
97				-0.10					0.18		0.13						0.20			
105	0.16	-0.20				0.48		0.05						0.27	0.22	0.13	-0.23	-0.31		
122	-0.07			0.45			0.25								0.72	0.03				-0.07
152	-0.04		0.23	0.23			0.15		-0.08		-0.08			0.41			0.13	0.14	-0.26	0.18
182			-0.66		-0.93		-0.71				0.77		0.37	0.43			2.29	3.10	0.75	0.38
183				-0.11								0.10		0.05		0.07		0.26	0.17	-0.08
198	-0.04				0.15											0.08	0.22		0.19	
272			0.08														0.16	0.16		0.30
303														-0.22	0.52	0.07	-0.43		1.92	

We assessed the fitness of the GLM models by assessing the normality of the residuals. If the model is reliable, the residuals should be Normally distributed, as demonstrated by the histogram of the standardized residuals in (**Figure 8**) and a Normal probability plot (P-P plot) in (**Figure 9**).

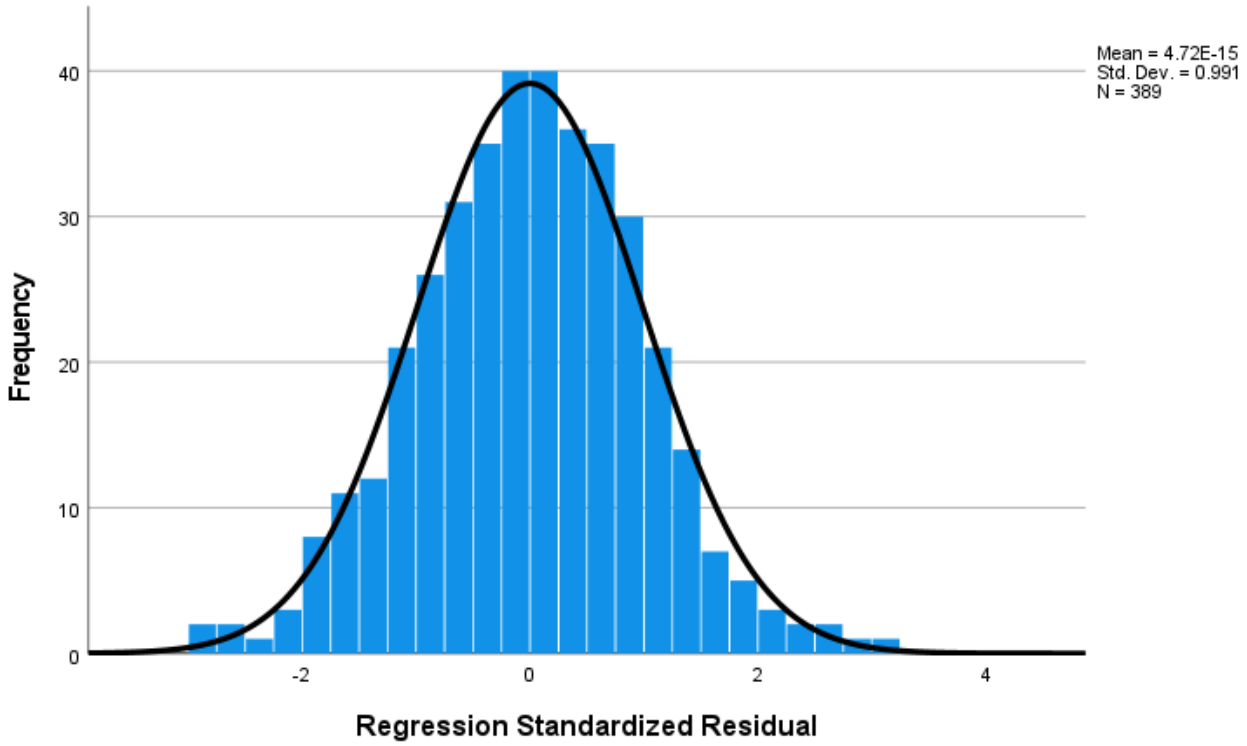

Figure 8. Histogram of standardized residuals of the m/z 105 from the training set model including an intercept (n = 389) compared to a Normal distribution.

Figure 8 shows a histogram of the standardized residuals using linear regression with an intercept for the training set of *m/z* 105. The black line represents a normal distribution. This plot shows that the residuals are normally distributed around zero, which means that the residuals are random and there is no variance left that could be explained.

Figure 9 displays the normal probability plot of the residuals of the training set model that includes a constant for *m/z* 105. This further demonstrates the normality of the error terms.

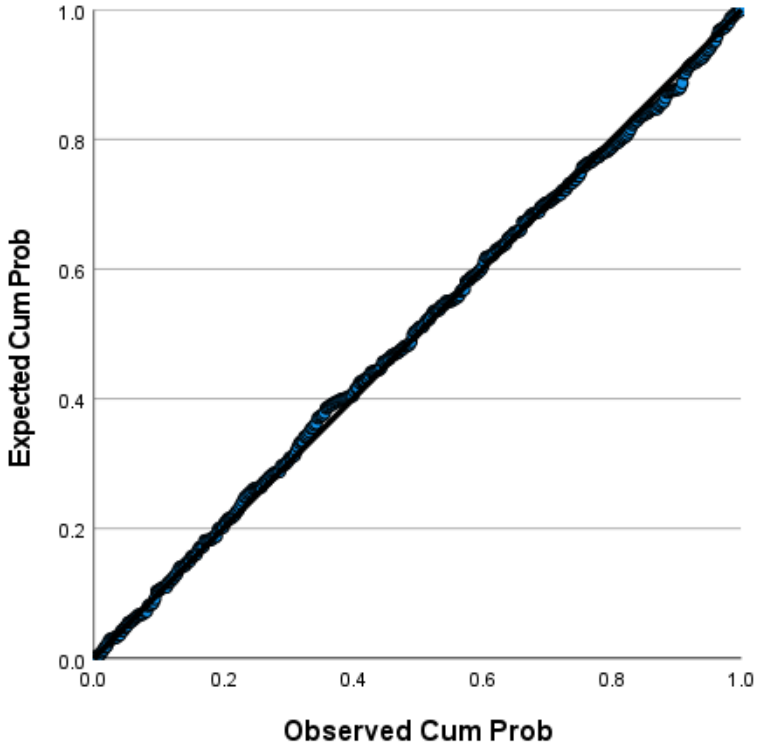

Figure 9. Normal probability plot showing the cumulative frequency of the distribution of the standardized residuals of the training set model including a constant (n=389) for *m/z* 105 compared to the normal probability graph scale.

One of the advantages of the EASI approach is the potential for extrapolation between spectra from different labs, as demonstrated in our first publication.[32] This potential can be visualized in **Figure 10**.

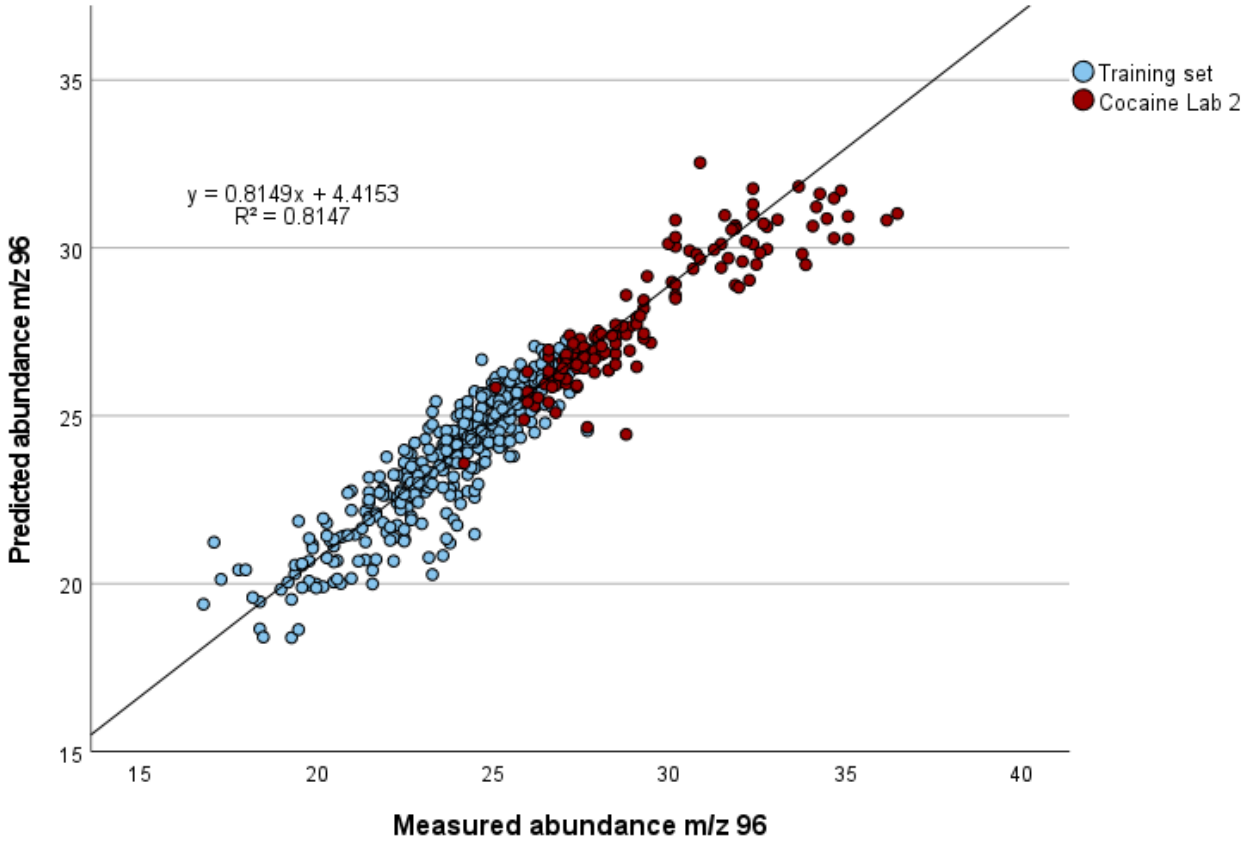

Figure 10. Scatter plot of measure versus predicted abundance of *m/z* 96 using EASI with a constant. The training set data (in blue) was collected at WVU Department of Forensic and Investigative Science (N = 389), whereas the Lab 2 data (in red) was curtesy of Benny Lum at Broward Sheriff's Office Crime Laboratory (N = 132).

For **Figure 10**, the replicate data from the full factorial design of experiments was used as the training set to build the GLM model for *m/z* 96 with a constant. The model was then applied to all the training set data and 132 cocaine spectra from a different laboratory. Details of the data acquisition for the data from Broward Sheriff's Office Crime Laboratory are provided elsewhere.[31,32] The mean measured abundance for *m/z* 96 for the two data sets are significantly different (two-tailed t-test, α=0.05); the training set has a mean abundance of 23.9% (s.d. = 2.1) and the test set from lab 2 has a mean of 29.4% (s.d. = 2.7). However, although the variance of the data from Lab 2 falls outside the variance of the training set range, the linear regression based on the training data can be extrapolated to effectively predict the abundances in Lab 2. In other words, the ability to extrapolate to Lab 2 the trend from the training set demonstrates the potential of EASI

in comparing inter-laboratory and inter-instrument data. Naïve, non-expert algorithms that lack a fundamental basis of operation like RRKM theory would not be able to predict that linear behavior could be extrapolated between instruments.

4.2.1 Penalization for the EASI without a constant

The exclusion of a constant in the multivariate linear modeling approach can be rationalized by the following thought experiment: If 19 of the 20 most abundant ions in a spectrum of a suspected drug measure 0%, the 20^{th} ion abundance should also be zero. There would be no reason to assume that the 20^{th} peak would register any signal. In contrast, the inclusion of an intercept in the GLM models simply allows the resultant model to better fit the data, regardless of chemical expectations. However, GLM modeling with and without a constant suffers one potential problem when known negatives have multiple ion abundances at or near 0%. Ion abundances of zero for known negatives make sense because we are only looking at the most abundant ions of cocaine, which might not include abundant ions for other substances. However, when each coefficient in the GLM is multiplied by ~0% abundance, the entire term (predicted dependent value) becomes ~zero. Therefore, any linear combination of zeros as the independent variables results in a predicted abundance of zero. This means that a known negative with all 20 measured ion abundances at or close to zero will have near-perfect predictions and therefore appear to fit the model well. From a mathematical perspective, the small residual error between the measured and known values of ~0% would mean that a known negative would appear to be a positive. **Figure 11** exemplifies the rise of false positives as a direct result of this problem. In this case, known negatives include methamphetamine, fentanyl, heroin and hydromorphone, which share almost no spectral overlap. Almost all the measured values are close to zero, so there are hundreds of known negative spectra provide that provide mean absolute residuals close to zero.

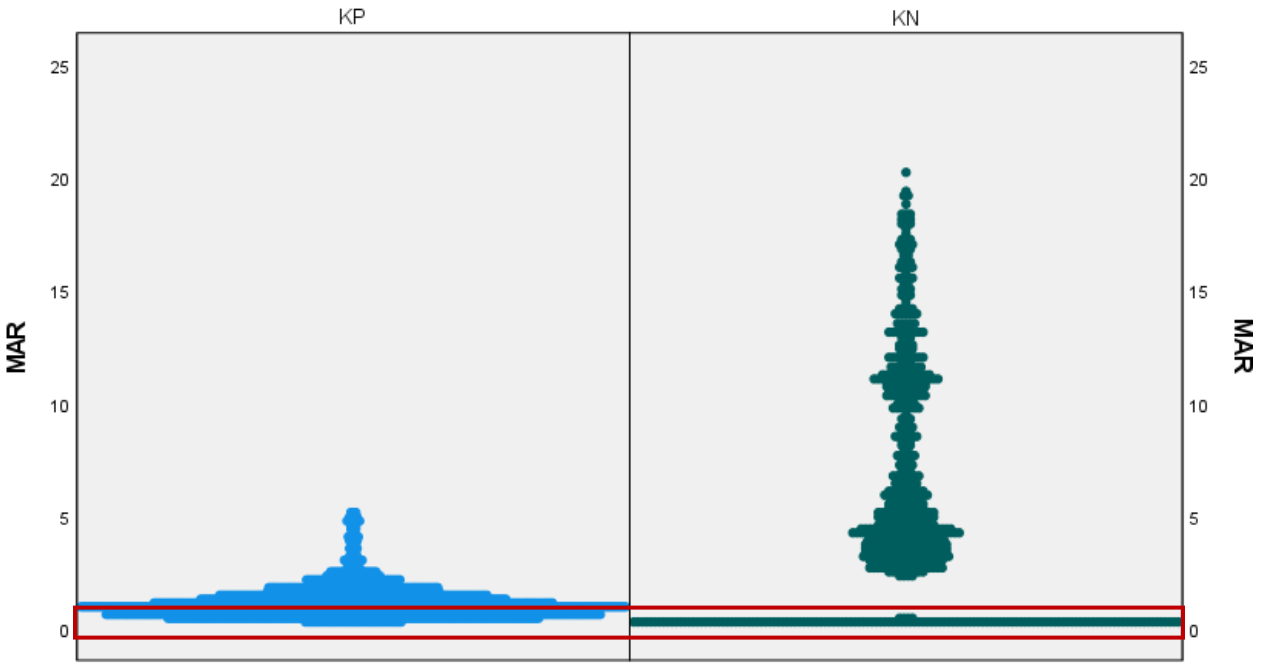

Figure 11. Population pyramid of mean absolute residuals (MAR) of known positives (N = 1478) and known negatives (N = 721) using EASI without a constant.

To avoid false positive classifications due to abundances at or close to zero, each measured ion abundance equal to zero was penalized with a value of 50%. This penalization included known positives and known negatives to maintain consistency. When each penalty of 50% is averaged across 19 predictions, the mean absolute residual is increased by 2.63% per penalty. This is usually enough to make a known negative have a significantly larger MAR than known positives.

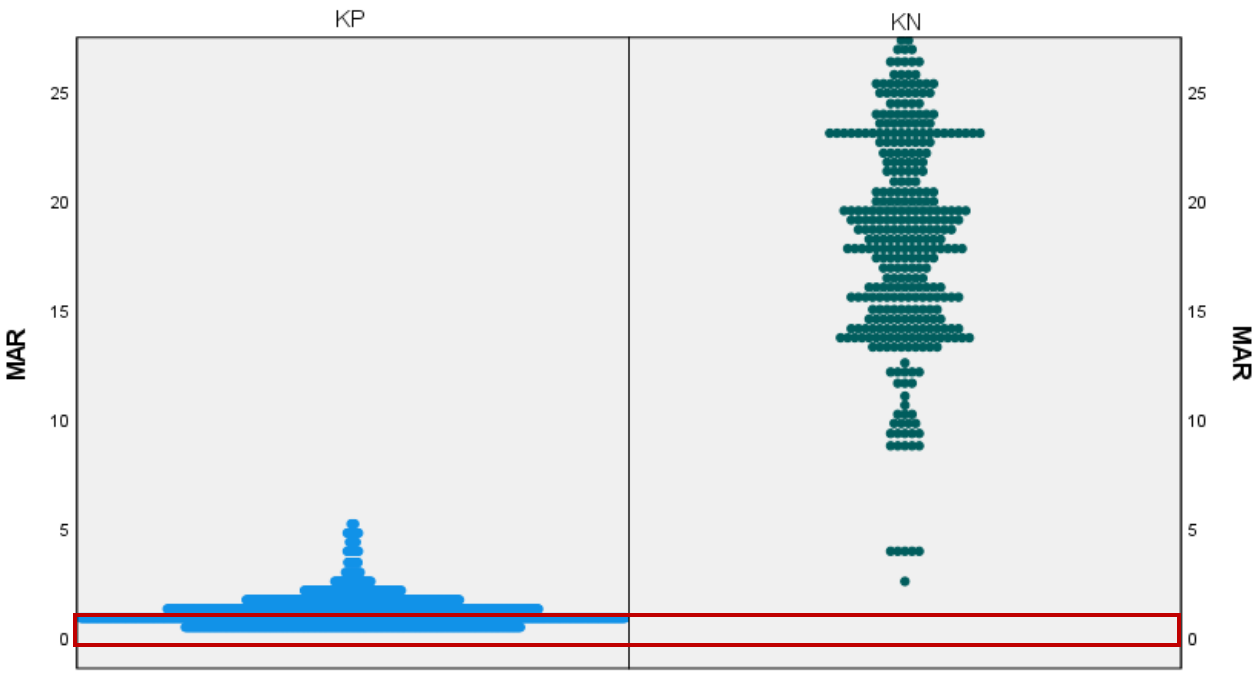

Figure 12. Population pyramid of mean absolute residuals (MAR) of known positives (N =1478) and known negatives (N=721) using EASI without an intercept in the models and with a penalization for abundances of zero.

Moving forward, EASI models without a constant and employing a penalty will be referred to as EASI without a constant (EASI WO).

4.2.2 Mean absolute residuals calculations and graphs

The mean absolute residual is a common measure by which to assess the residuals in multivariate modeling. Consequently, the MARs were calculated as described before from the 19 residuals for the 19 predictions for each spectrum. **Table 14** presents a summary of minimum and maximum mean absolute residuals to highlight the possibility of overlap between the distributions of known positives and known negatives.

Table 14. Minimum and maximum mean absolute residuals (MAR, %) by compound and model. Bold values are the largest MARs for known positives. Underlined values are the smallest MARs for known negatives that overlap with the distribution of known positives.

	EASI		**EASI WO**		**Consensus**	
	Min	Max	Min	Max	Min	Max
KPs (Training set)	0.22	1.80	0.24	1.90	0.67	5.76
KPs (Test set)	0.22	**5.07**	0.29	**5.22**	0.89	**14.40**
Pseudoallococaine	2.74	4.37	4.03	13.68	4.12	4.58
Pseudococaine	2.60	4.59	2.49	22.28	4.11	10.58
Allococaine	3.56	3.56	8.66	8.66	6.40	6.40
Ecgonine methyl ester	7.92	10.80	24.32	43.95	13.18	15.61
Fentanyl	5.26	8.66	11.95	41.20	16.01	20.36
Heroin	4.65	13.12	8.69	37.54	16.98	20.00
Hydromorphone	6.95	18.24	13.87	39.90	15.29	19.98
Methamphetamine	4.80	5.17	35.06	49.60	21.95	22.19

The mean absolute residuals for the training set were the smallest for all three models. This was expected because the training set was used to build the model, so there is some bias in favor of good predictions for this set. However, the range of MARs is considerably wider for the consensus model. For the consensus model, KPs from the test set have a MAR as large as 14.40%. On the other hand, while the ranges of MARs for both EASI models also grew larger for the test set of KPs relative to the training set of KPs, the maximum only reached 5.07% and 5.22%, respectively, for EASI and EASI WO.

For the cocaine diastereomers (pseudococaine, allococaine and pseudoallococaine), EASI and EASI WO yields non-overlapping ranges between the diastereomers and the training set. However, there is overlap between the cocaine test set and its diastereomers. The overlap in distributions means that a binary classifier based on MARs could not be 100% accurate. Depending on the threshold, a binary classifier would either result in false positives (diastereomers being classified as cocaine), false negatives (cocaine being classified as one of the diastereomers), or

both, depending on the threshold. For the consensus approach, the ranges for pseudococaine and pseudoallococaine overlap with the those of training set so the consensus approach would be less effective at correctly predicting binary classification for the diastereomers relative to the training set. These trends already show that EASI makes more accurate spectral predictions for KP's than the consensus approach, so EASI has significantly greater potential for successful binary classification than the consensus approach.

Table 14 also shows that for EASI there are two more compounds that have overlapping ranges with the test set of KP's: heroin and methamphetamine. For the EASI WO there are no other known negatives that present overlapping ranges that could lead to false positives or false negatives. Lastly, the consensus model shows overlapping ranges between the cocaine and ecgonine methyl ester.

The figure below (**Figure 13**) shows population pyramids from EASI using the mean absolute residuals separated by known positives, and various known negatives. **Figure 13** shows a tight cluster with a narrow tail for the KP's, which range from 0.22 to 5.07, according to Table 14. The distribution of MARs for some KN's are very wide, like hydromorphone, which spans from 6.95 to 18.24. The distribution in MARs for other KNs are much tighter, like methamphetamine, which spans from 4.80 to 5.17. This figure helps visualize the overlap mentioned before between the known positive cocaine spectra in the test set and the known negative spectra of cocaine diastereomers, heroin and methamphetamine. However, **Figure 13** shows that the overlap between the known positives and the cocaine diastereomers is more extensive than the overlap between heroin and methamphetamine. As stated before, these overlapping distributions create a problem when trying to set a threshold that eliminates false positives and false negatives.

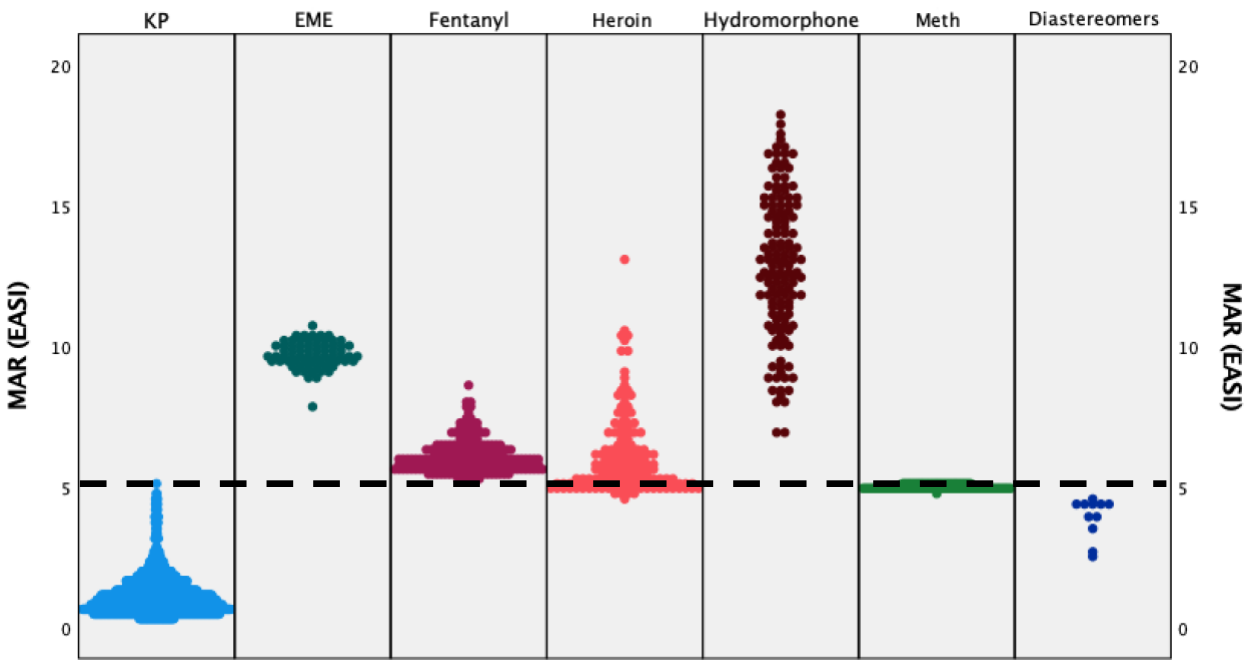

Figure 13. Population pyramid from EASI using the mean absolute residuals. N_{KP}=1478, N_{EME}=69, $N_{fentanyl}$=216, N_{heroin}=158, $N_{hydromorphone}$=134, N_{meth}=133, $N_{diastereomers}$=11.

Figure 14 shows a close-up view of the EASI population pyramid by all the known positives (left) and known negatives (right). This figure allows us to see that the overlap between known positives, heroin and methamphetamine ranges is mainly due to a couple of cocaine datapoints and not the bulk of known positives. In general, the distributions of KPs and KNs are very well separated. Of course, if more spectra of diastereomers were included, the distributions would overlap much more. Unfortunately, additional spectra of cocaine diastereomers could not be obtained because there are no commercial vendors available. Several vendors list cocaine diastereomers in their inventories, but when we requested quotes, they reported that they could not validate the isomeric purity unless we paid a substantial fee or completed a bulk order.

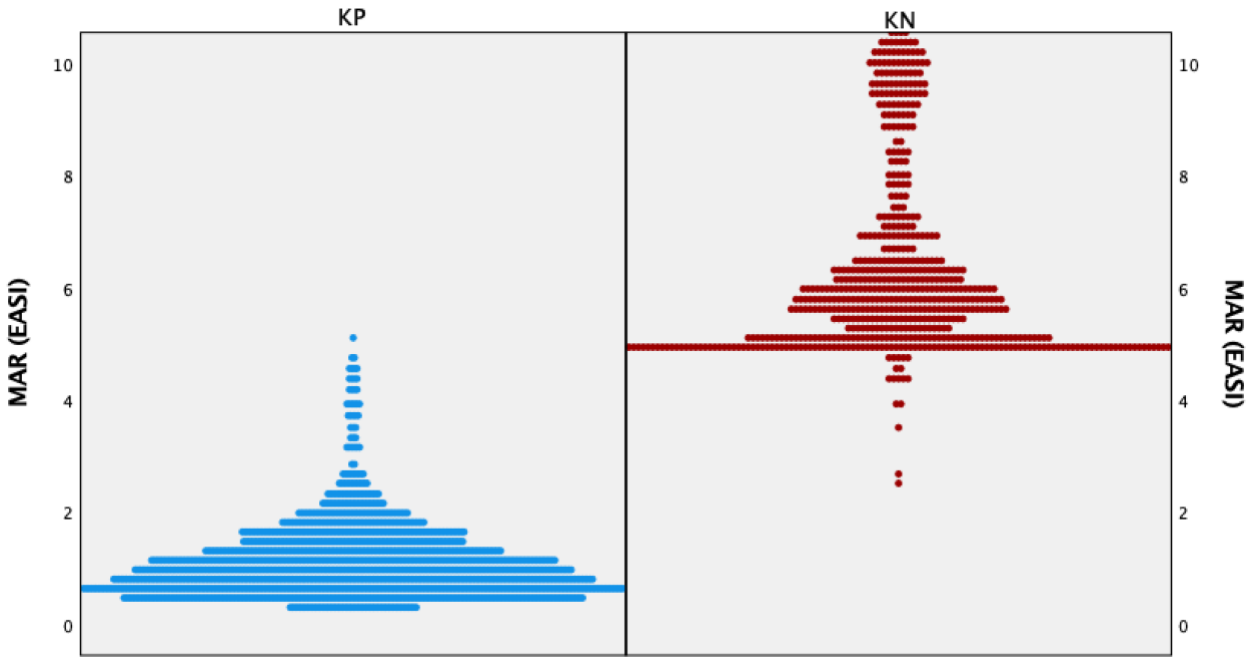

Figure 14. Close-up view of the population pyramid from the EASI using the mean absolute residuals. The left distribution (in blue) shows the known positives (N=1478), and the right distribution (in red) represents the known negatives (N=721).

Figure 15 presents the EASI WO population pyramid of the MARs by KPs (to the left in blue) and KNs classified by compounds. The KP's range (0.24-5.22) is similar to that of EASI with constants in the models. However, one of the main differences between EASI and EASI WO is the wider range of MARs for the known negatives as a result of the penalization process. This trend is evident especially for the methamphetamine, whose range increased from 4.80-5.17 for EASI to 35.06-49.60 for EASI WO, and ecgonine methyl ester, whose range increased from 7.92-10.80 for EASI to 24.32-43.95 for EASI WO. For EASI WO, the diastereomers are the only column that provide any overlap with the KPs.

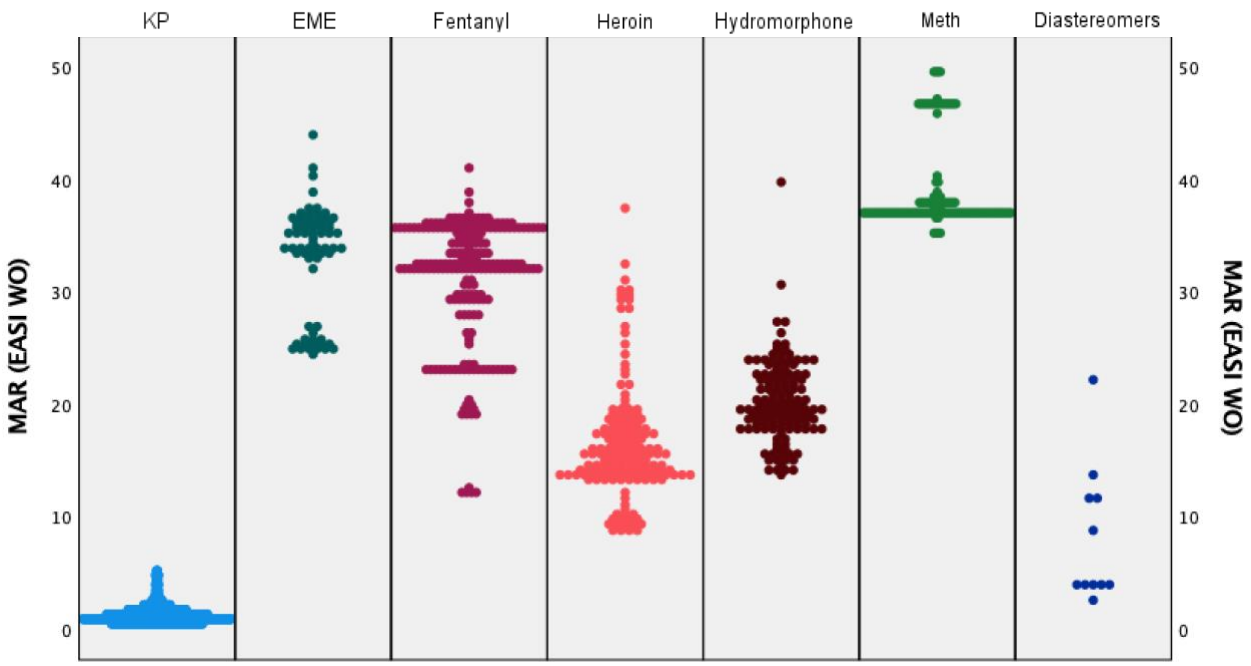

Figure 15. Population pyramid from EASI WO using the mean absolute residuals N_{KP}=1478, N_{EME}=69, $N_{fentanyl}$=216, N_{heroin}=158, $N_{hydromorphone}$=134, N_{meth}=133, $N_{diastereomers}$=11.

Figure 16 shows population pyramids for the MARs using the consensus approach. Compared to the EASI and EASI WO models in **Figure 13** and **Figure 15**, the consensus approach in **Figure 16** shows a much wider spread in the MARs for the KPs. This wider range causes significant overlap with the cocaine diastereomers and one KP overlaps with the distribution of MARs of ecgonine methyl ester.

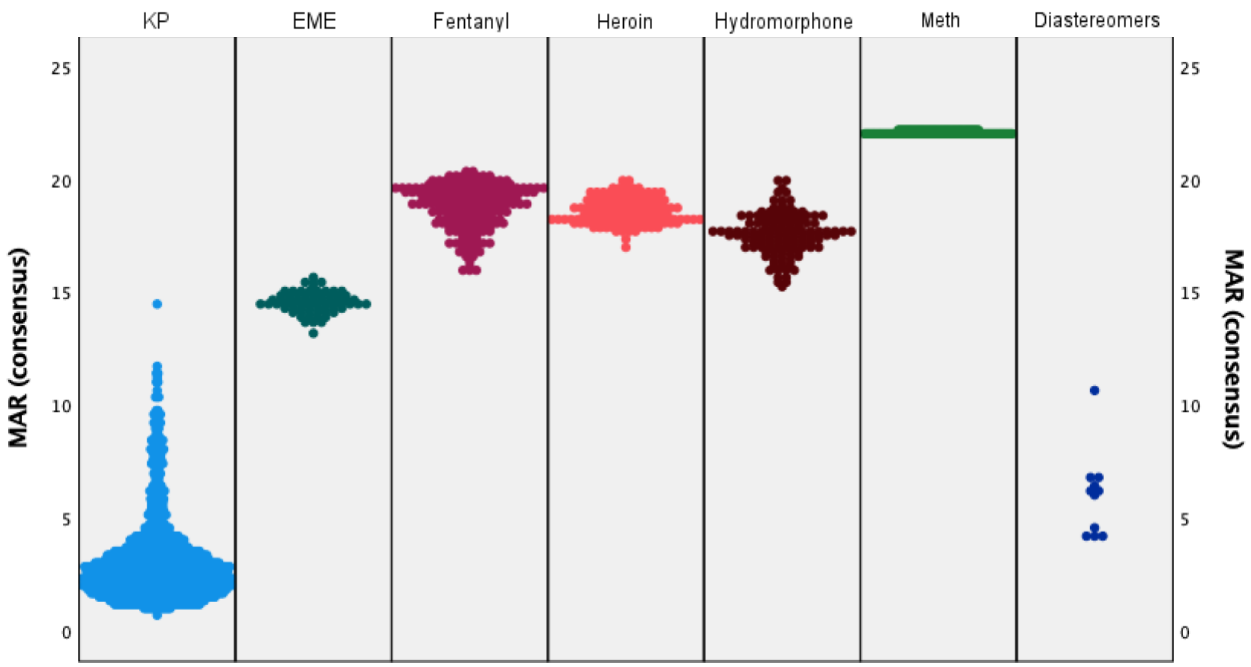

Figure 16. Population pyramid from the consensus model using the mean absolute residuals. N_{KP}=1478, N_{EME}=69, $N_{fentanyl}$=216, N_{heroin}=158, $N_{hydromorphone}$=134, N_{meth}=133, $N_{diastereomers}$=11.

Figure 17 shows ROC curves generated from the mean absolute residuals of EASI, EASI WO, and the consensus model using all known positives and all known negatives. The curves for all three models look similar because the known negatives are dominated by spectra that are easy to resolve from cocaine, like hydromorphone, fentanyl, and heroin.

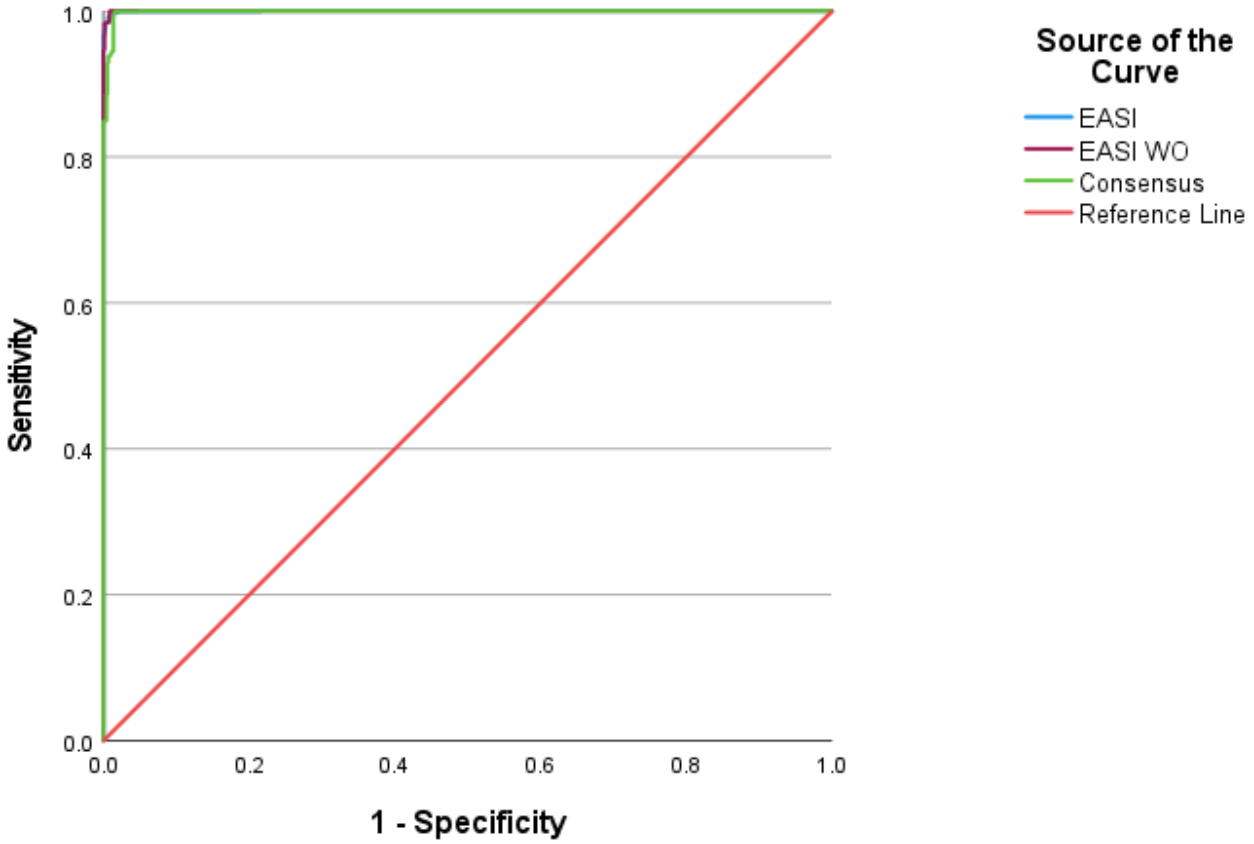

Figure 17. ROC curves generated using the mean absolute residuals of all known positives (N=1478) and negatives (N=721).

Table 15 presents the AUCs for the ROC curve in **Figure 17**. These AUCs further demonstrate the similarities in results for the three models when analyzing all known positives and known negatives. Again, the classification rates are dominated by spectra other than the diastereomers.

Table 15. The AUCs generated using the mean absolute residuals and all known negatives and positives.

Model	AUC
EASI	0.9997
EASI WO	0.9998
Consensus	0.9987

EASI and EASI WO had very high and similar AUCs which demonstrate the effectiveness of the models when coupled with MARs as a binary classifier. However, the consensus approach

also had a high AUC of 0.9987, which is comparable to the EASI and EASI WO. These values show that the differences between the three models are not highlighted when analyzing all the known negatives because most of the known negatives are easy to distinguish.

To highlight the model differences, **Figure 18** shows the ROC curve generated from the mean absolute residuals of the EASI, EASI WO, and the consensus model using all known positives and only the cocaine diastereomers as known negatives.

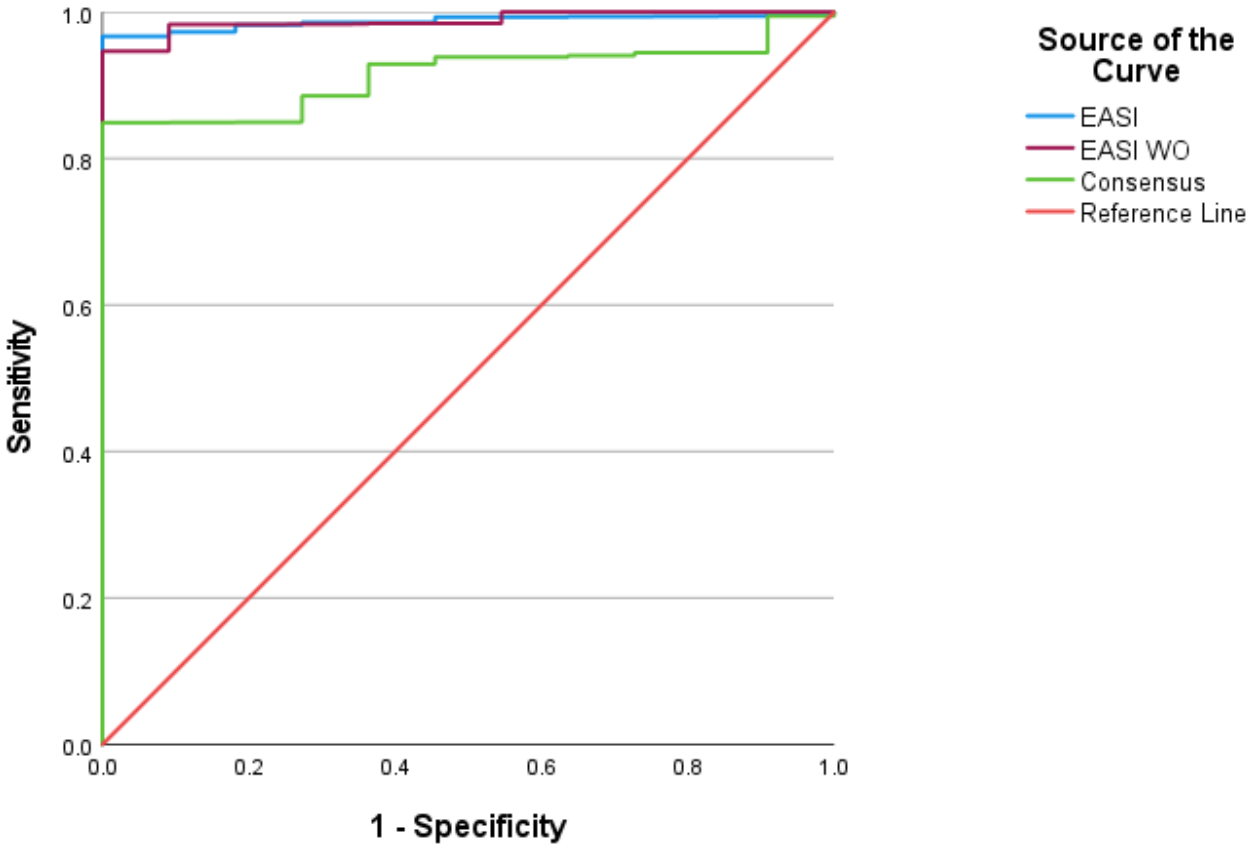

Figure 18. ROC curve generated using the mean absolute residuals of all known positives (N=1478) and, pseudococaine, allococaine and pseudoallococaine as known negatives (N=11).

Taking only the cocaine diastereomers as known negatives allows for the examination of the models when faced with more structurally similar and spectrally similar compounds. This similarity in structure poses a greater challenge for the models when trying to classify them as cocaine or not cocaine.

Table 16. The AUCs generated using the mean absolute residuals and all known positives and the cocaine diastereomers as the known negatives.

Model	AUC
EASI	0.9873
EASI WO	0.9878
Consensus	0.9149

Table 16 shows that the EASI WO has the highest AUC, with a 98.8% chance of correctly ranking a cocaine sample as cocaine. The EASI is very close with a discriminatory chance of 98.7%, whereas the consensus approach is only 91.5% effective at distinguishing cocaine from its diastereomers. The 0.1% difference represents 1 extra false identification between the two EASI models. The difference between the consensus model and EASI WO model is 109 false identifications, most of which are false positives of cocaine (there are only 11 known negatives available).

4.2.3 Euclidean distance calculations and graphs

The Euclidean distance is another measure used to assess the residuals in multivariate modeling, and it is the shortest straight-line distance between two points in multidimensional space. The Euclidean distances were calculated as described before from the 19 residuals for the 19 predictions for each spectrum. **Table 17** presents a summary of minimum and maximum Euclidean distances.

Table 17. Minimum and maximum Euclidean distances by compound and model. Bold values are the largest Euclidean distances for known positives. Underlined values are the smallest Euclidean distances for known negatives that overlap with the distribution of known positives.

	EASI		**EASI WO**		**Consensus**	
	Min	Max	Min	Max	Min	Max
KPs (training set)	1.28	12.15	1.66	15.11	4.39	39.95
KPs (test set)	1.31	**39.23**	1.74	**39.79**	5.12	**96.63**
Pseudoallococaine	17.16	26.10	24.84	101.67	28.03	32.12
Pseudococaine	14.41	27.54	14.17	165.91	28.01	64.71
Allococaine	22.10	22.10	65.39	65.39	44.23	44.23
Ecgonine methyl ester	46.88	65.89	151.65	395.65	86.77	98.67
Fentanyl	35.33	57.07	92.37	387.52	124.25	136.53
Heroin	30.99	83.16	62.70	299.39	120.42	136.04
Hydromorphone	41.83	119.82	88.37	274.27	96.51	138.34
Methamphetamine	31.45	32.83	314.48	397.48	142.83	143.56

As with the mean absolute residuals, the training provided the smallest upper limit for all three models. The maximum value for the known positives for the EASI and EASI WO were very similar at 39.23 and 39.79, respectively. Similar to the trend described for MAR's, the consensus Euclidean distance range for known positives (5.12-96.63) was considerably larger than EASI (1.31-39.23) and EASI WO (1.74-39.79). This larger range can pose a problem when trying to use the Euclidean distance with the consensus model to classify query samples as cocaine or not cocaine.

For the cocaine diastereomers, both EASI and the consensus model have completely overlapped ranges that would not allow the differentiation between cocaine and its diastereomers. On the other hand, the EASI WO has the potential to distinguish the diastereomers because the ranges do not completely overlap. For pseudococaine and pseudoallococaine, several Euclidean distances are larger than the upper limit of the KP's range, and for allococaine there is no overlap at all.

For the rest of the known negatives, **Table 17** shows two partial overlaps using EASI between the KPs and fentanyl (35.33-57.07) and heroin (30.99-83.16). This means that the ranges for these known negatives are not entirely within the KP's range, but overlap a little. In contrast, all the methamphetamine spectra have Euclidean distances (31.45-32.83) that fall within the range of KP cocaine spectra. The consensus model has two partially overlapping ranges with the KPs: ecgonine methyl ester (86.77-98.67) and hydromorphone (96.51-138.34). The EASI WO model was the only model with no overlaps besides pseudococaine and pseudoallococaine.

The figure below (**Figure 19**) shows population pyramids for EASI using the Euclidean distances separated by known positives and a variety of known negatives.

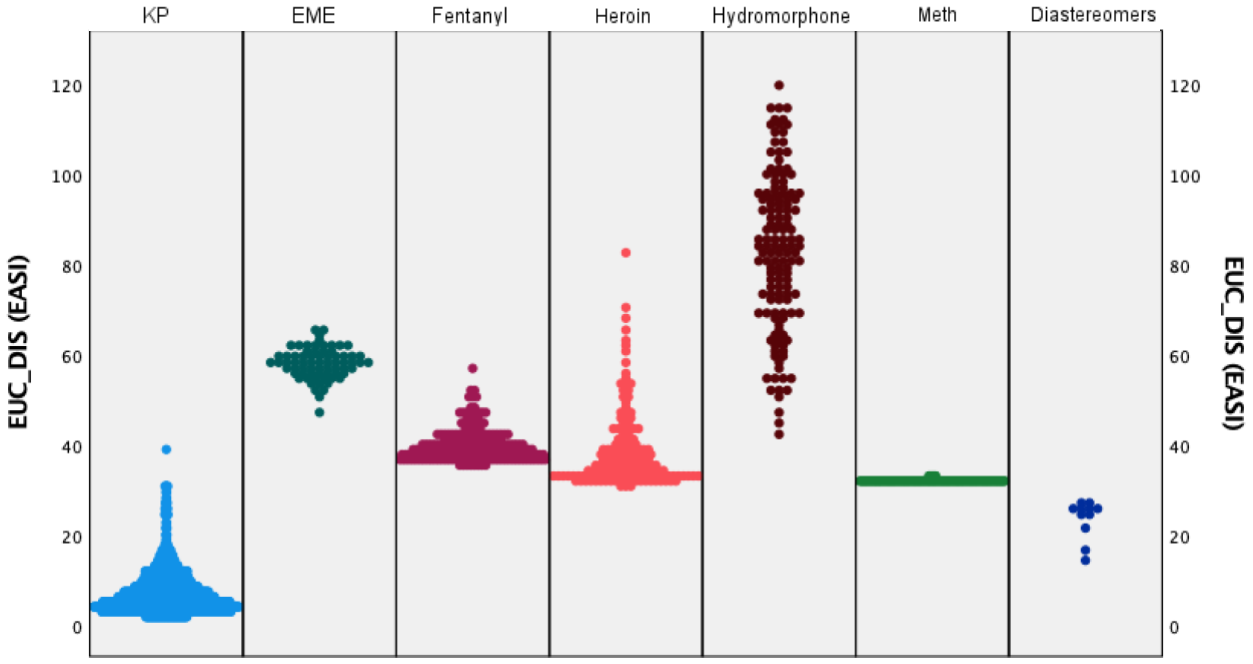

Figure 19. Population pyramid from the EASI using the Euclidean distances. N_{KP} =1478, N_{EME}=69, $N_{fentanyl}$=216, N_{heroin}=158, $N_{hydromorphone}$=134, N_{meth}=133, $N_{diastereomers}$=11.

Figure 19 shows the overlapping distributions discussed above. Additionally, this figure shows where the bulk of the data is within each category. The distribution of known positives shows a narrow tail with a single datapoint farther away from the main cluster. This datapoint expands the spread of the known positives to the point of overlapping with methamphetamine's

tight distribution and some samples of heroin and fentanyl. The figure also shows the overlap of the cocaine diastereomers with the known positives.

Figure 20 shows a summary population pyramid split by all known positives and known negatives to highlight the overlap as a binary classifier. The datapoint around 40 in the distribution of known positives belongs to a NIST spectra that has a base peak at *m/z* 182. Although cocaine can have its base peak at either *m/z* 82 or 182, none of the collected data has the base peak at *m/z* 182. This difference in behavior results in a lack of spectra in the training set that could help minimize the spread in the distribution. In prior work cocaine spectra from two different crime labs each contained some replicates with base peaks at either *m/z* 82 or 182.[32,43]

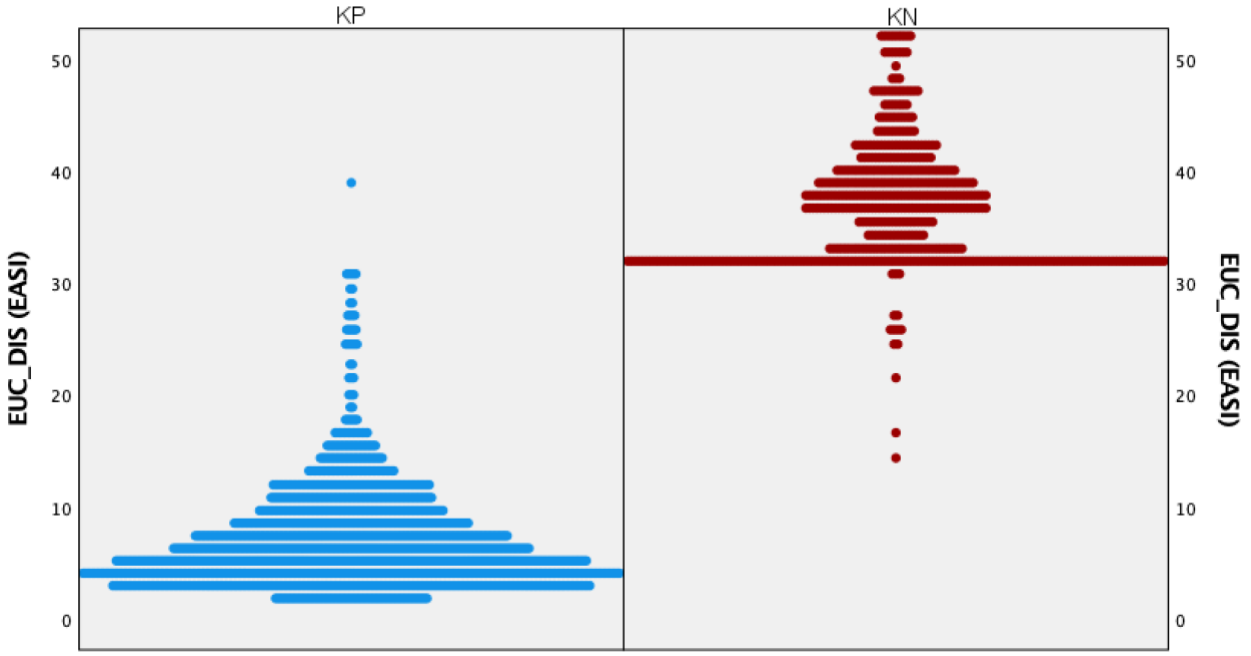

Figure 20. Close-up view of the population pyramid from the EASI using the Euclidean distances. The left distribution (in blue) shows the known positives (N=1478), and the right distribution (in red) represents the known negatives (N=721).

Figure 21 presents the population pyramid from the EASI WO using the Euclidean distances separated by known positives and a variety of known negatives.

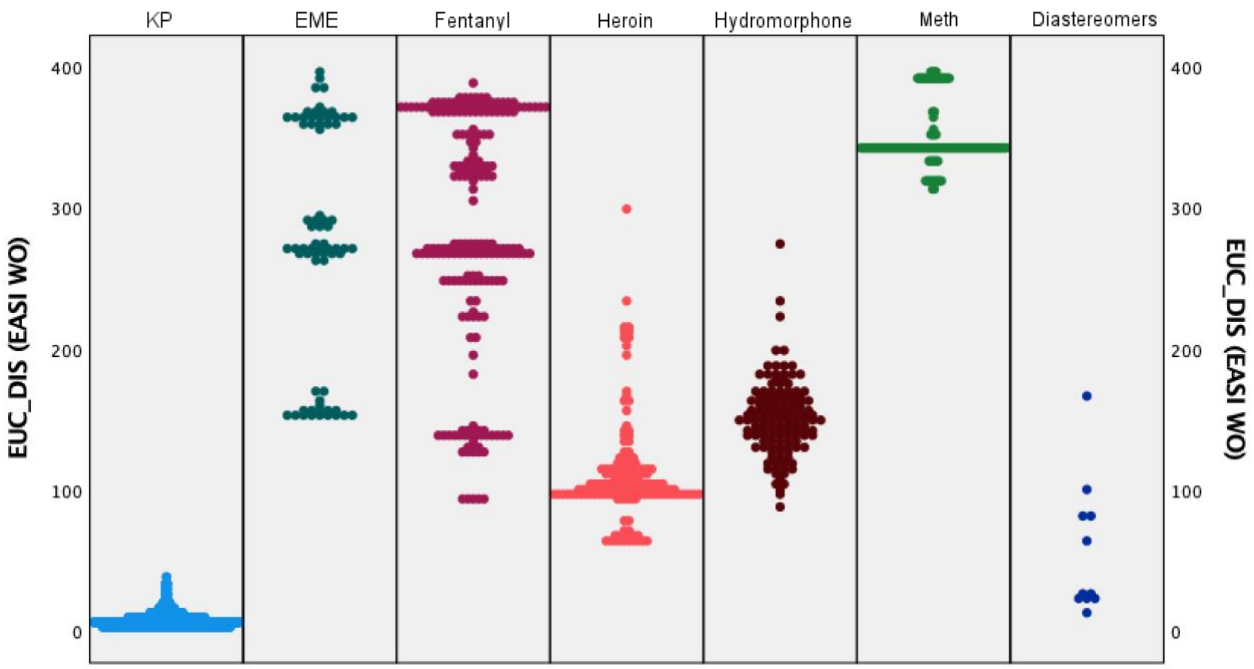

Figure 21. Population pyramid from EASI WO using the Euclidean distances. N_{KP}=1478, N_{EME}=69, $N_{fentanyl}$=216, N_{heroin}=158, $N_{hydromorphone}$=134, N_{meth}=133, $N_{diastereomers}$=11.

Figure 21 shows how the penalization process divided some compound's distributions depending on how many penalties were incorporated. For example, the first penalty of ~50% to one *m/z* channel adds 50% to the Euclidean distance. Two penalties of 50% provide a total penalty of 70% to the Euclidean distance, so the second penalty effectively adds an additional 20% to the first penalty. The effective magnitude of each additional penalty decreases, so that the 10th penalty (if there was one) would only add 8% to the growing Euclidean distance. Each additional penalty caused by a measured ion abundance of zero in a particular channel leads to a discrete jump in the Euclidean score. At least four such groups are observed for ecgonine methyl ester, and many more are observed for fentanyl and methamphetamine compared to the distributions they exhibited with EASI.

Figure 22 shows a close-up of the population pyramid divided only by known positives and known negatives.

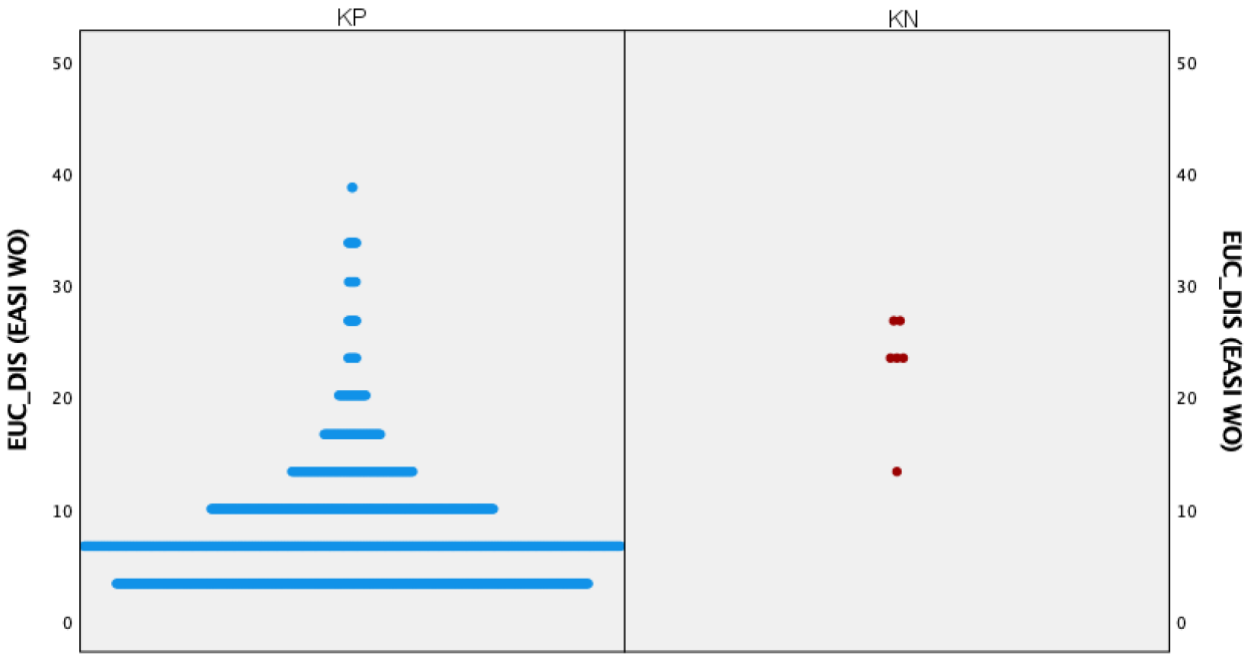

Figure 22. Close-up view of the population pyramid from EASI WO using the Euclidean distances. The left distribution (in blue) shows the known positives (N=1478), and the right distribution (in red) represents the known negatives (N=721).

Although the range of the known positives for EASI and EASI WO are similar (**Table 17**), their distributions are different. Whereas EASI had a single datapoint reaching a maximum Euclidean distance of 39.23, EASI WO has a tail with no visible outliers. However, like EASI, spectra in the tail of the distribution of known positives is due to spectra from the NIST archive with a base peak at *m/z* 182 instead of *m/z* 82. **Figure 22** also shows the minimal overlap between known positives and known negatives, with only six diastereomers falling within the range of known positives.

Figure 23 shows the population pyramid from the consensus approach using the Euclidean distances separated by known positives and a variety of known negatives.

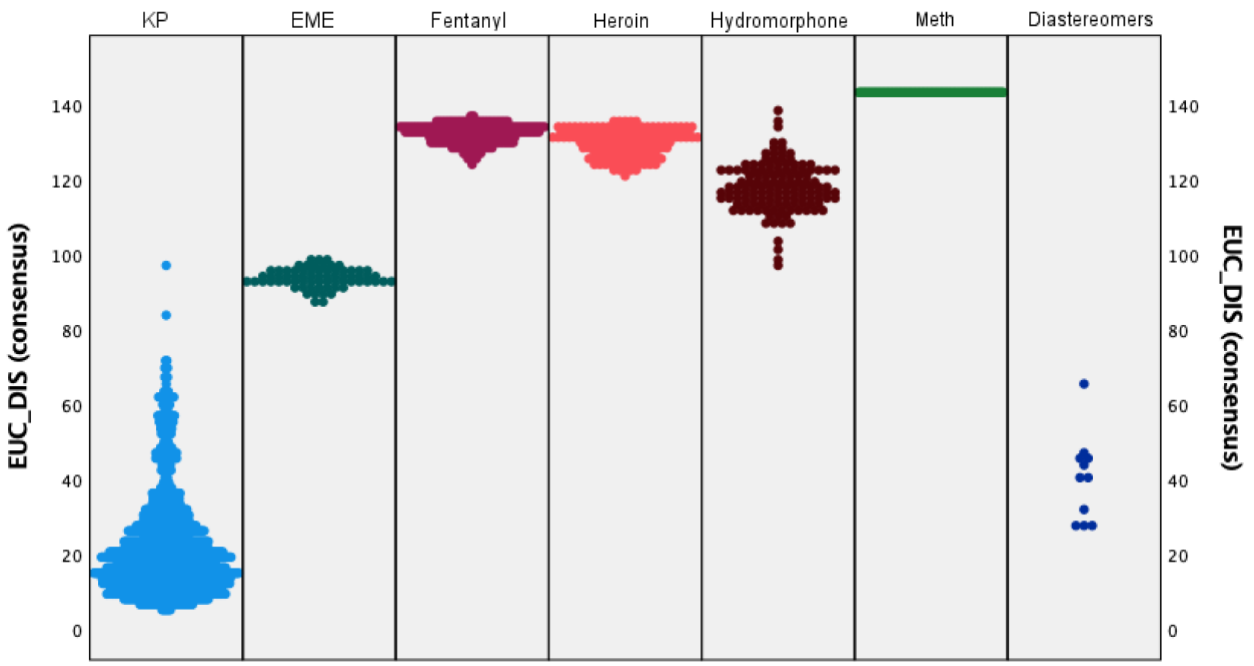

Figure 23. Population pyramid from the consensus model using the Euclidean distances. N_{KP}=1478, N_{EME}=69, $N_{fentanyl}$=216, N_{heroin}=158, $N_{hydromorphone}$=134, N_{meth}=133, $N_{diastereomers}$=11.

Here, the distribution of known positives is much wider that either EASI or EASI WO, with a maximum value of 96.6. Two of the cocaine spectra are obvious outliers. Again, these two datapoints are due to differences in base peaks and can cause an increase in false positives, specifically with ecgonine methyl ester, when using this metric as a binary classifier. **Figure 24** shows a close-up of the population pyramids separated by known positives and known negatives.

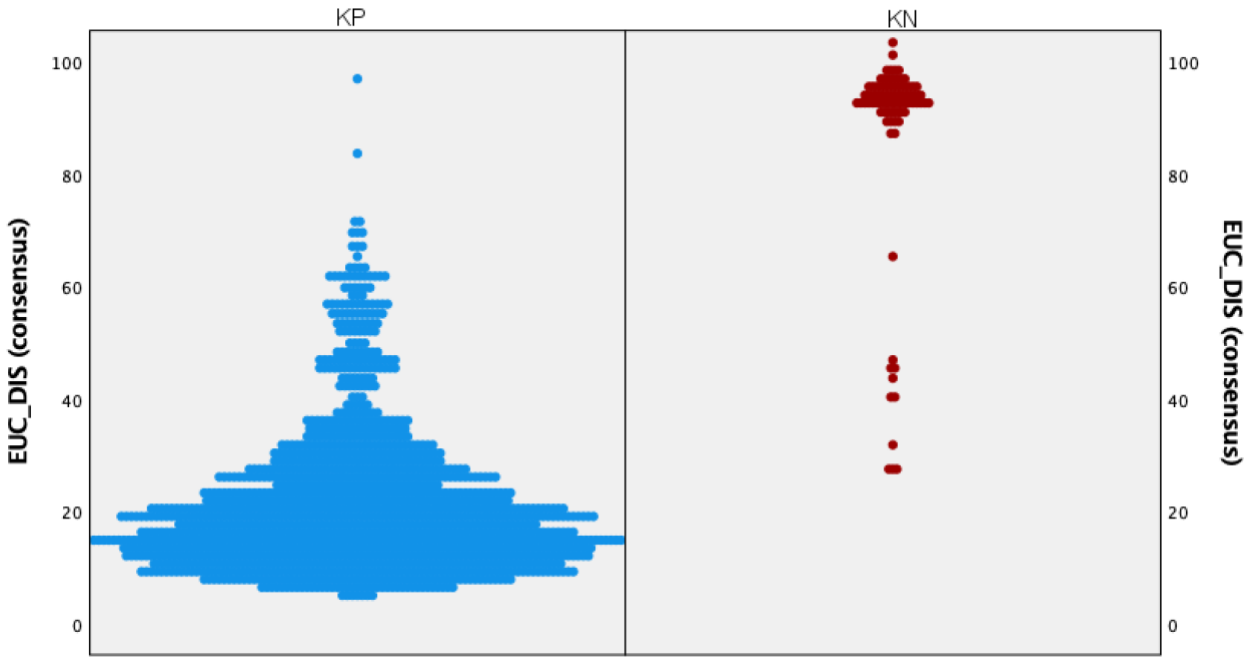

Figure 24. Close-up view of the population pyramid from the consensus model using the Euclidean distances. The left distribution (in blue) shows the known positives (N=1478), and the right distribution (in red) represents the known negatives (N=721).

Figure 24 above shows the known negatives that could be classified as cocaine if we chose a threshold with zero false negatives. This figure also shows that even if the two outliers were removed, binary classification is likely to result in many incorrect classifications.

Figure 25 shows the ROC curve generated from the Euclidean distances of EASI, EASI WO and, the consensus model using all known positives and known negatives. The differences are not visible because most of the compounds are spectrally distinct from cocaine and therefore correctly classified. This is further proved by the AUCs of the ROC curves shown in **Table 18**. The AUCs for all three models were close to 1, with the highest value being 0.9997 for EASI WO. For all three models, AUCs using Euclidean distances and MAR's were almost indistinguishable within models.

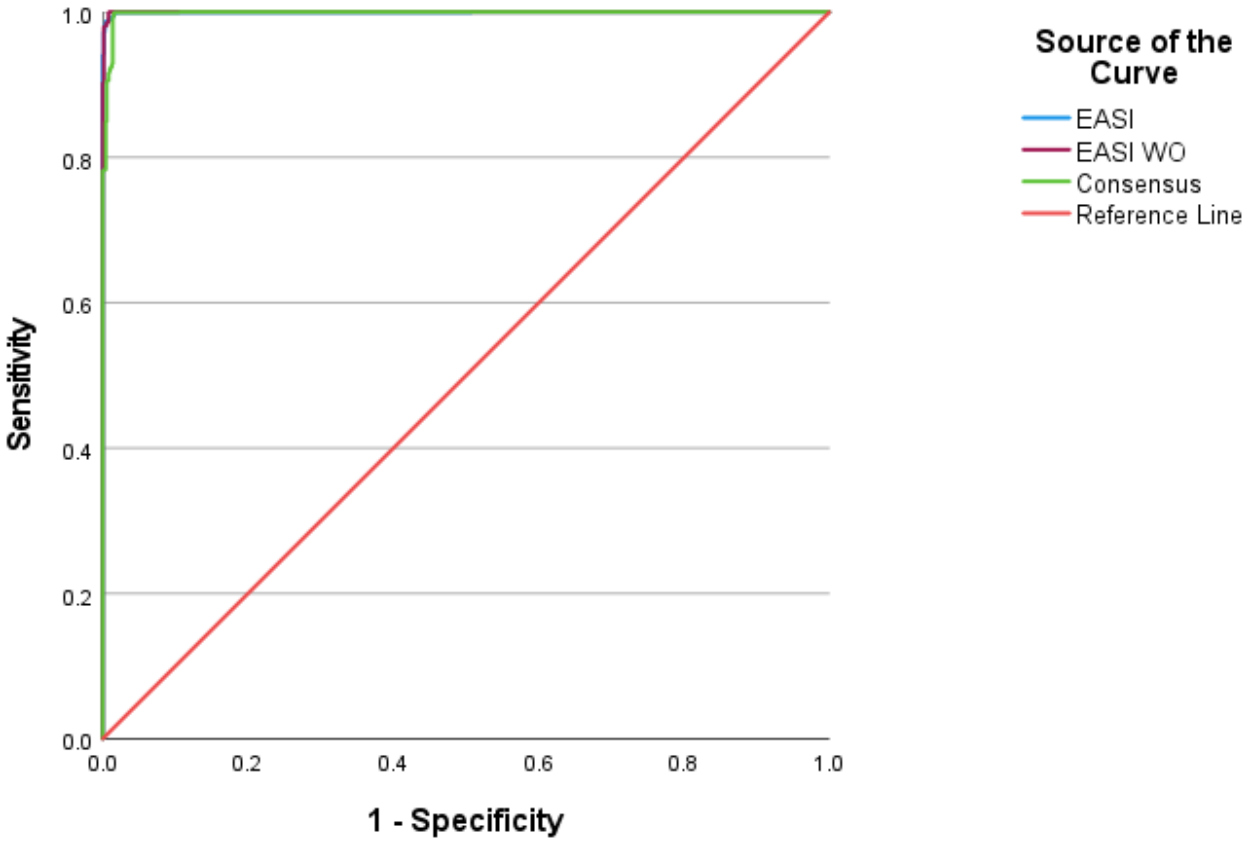

Figure 25. ROC curve generated using the Euclidean distances of all known positives (N=1478) and negatives (N=721).

Table 18. The AUCs generated using the Euclidean distances and all known negatives and positives.

Model	AUC
EASI	0.9994
EASI WO	0.9997
Consensus	0.9981

To better compare the performance of the models, new ROC curves were generated using the same known positives and only the cocaine diastereomers as known negatives, since they are the most difficult to distinguish from cocaine.

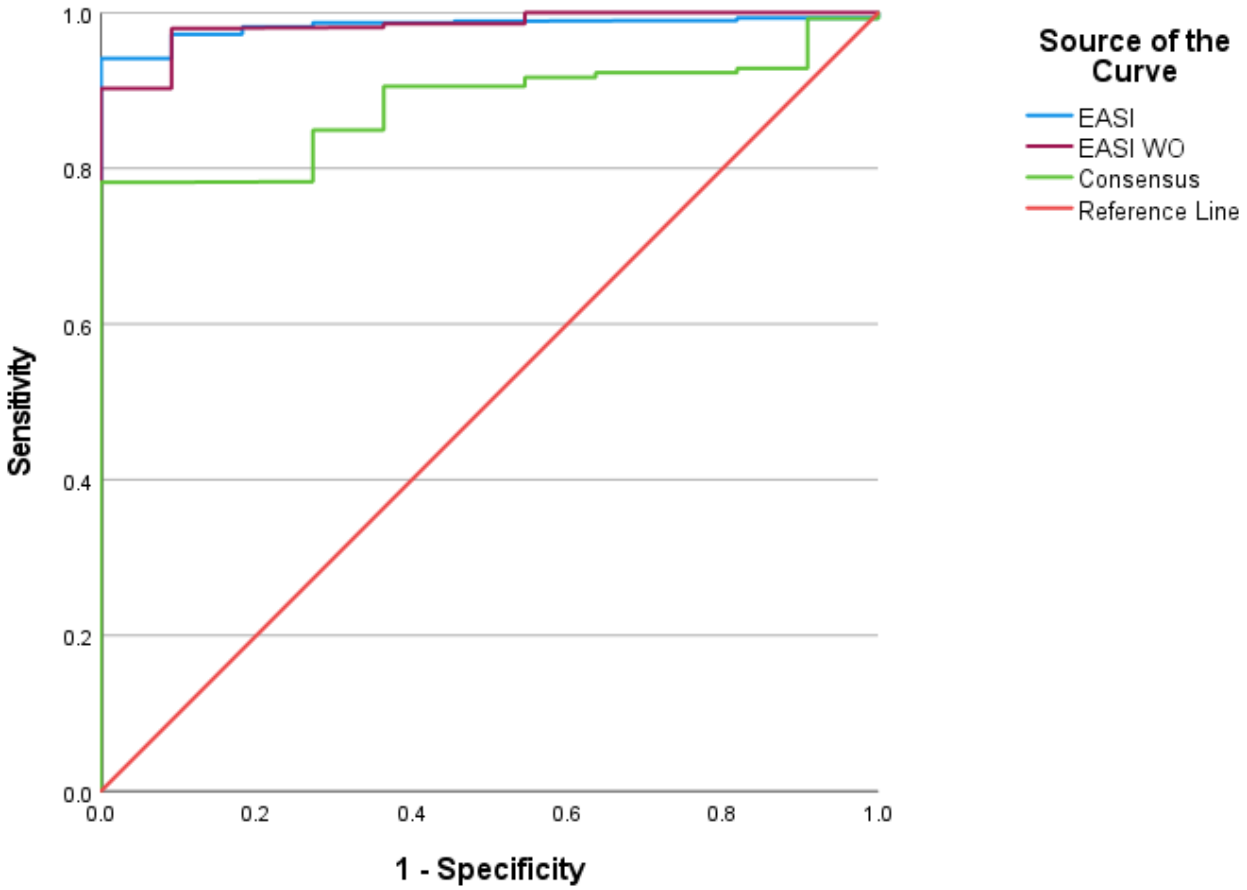

Figure 26. ROC curve generated using the Euclidean distances of all known positives (N=1478) and only pseudococaine, allococaine and pseudoallococaine as known negatives (N=11).

Figure 26 shows the ROC curves using the cocaine diastereomers as known negatives. Clearly, both EASI models offer superior performance relative to the consensus approach. **Table 19** shows that EASI and EASI WO have the same AUC, with a 98.3% chance of correctly identifying a sample, whereas the consensus model only has an 88.1% chance of correct identification to the binary groups.

Table 19. The AUCs generated using the Euclidean distances and all known positives and the cocaine diastereomers as the known negatives.

Model	AUC
EASI	0.9831
EASI WO	0.9832
Consensus	0.8809

4.2.4 PPMC calculations and graphs

The Pearson product-moment correlation (PPMC) coefficients were also calculated using the predicted and measured abundances for all 19 *m/z* values, as previously described (**Table 20**).

Table 20. Minimum and maximum PPMC coefficients by compound and model. Bold values are the smallest PPMCs for known positives. Underlined values are the largest PPMCs for known negatives that overlap with the distribution of known positives.

	EASI		**EASI WO**		**Consensus**	
	Min	Max	Min	Max	Min	Max
KPs (training set)	0.987	0.9999	0.994	0.9999	0.931	0.9991
KPs (test set)	**0.911**	0.9999	**0.949**	0.9999	**0.599**	0.9991
Pseudoallococaine	0.916	<u>0.964</u>	0.895	<u>0.973</u>	0.962	<u>0.969</u>
Pseudococaine	0.916	<u>0.990</u>	0.654	<u>0.993</u>	0.842	<u>0.969</u>
Allococaine	0.924	<u>0.924</u>	0.892	0.892	0.937	<u>0.937</u>
Ecgonine methyl ester	0.694	0.795	0.076	0.457	0.642	<u>0.704</u>
Fentanyl	0.058	0.369	-0.310	-0.089	0.177	0.237
Heroin	-0.176	0.195	-0.331	0.073	0.059	0.192
Hydromorphone	-0.049	0.573	-0.194	0.380	0.101	0.467
Methamphetamine	-0.318	-0.147	-0.368	-0.170	-0.004	0.102

Unlike the metrics before, where the overlap between known positives and known negatives was due to the known positives' maximum value and the known negatives' minimum value, for the PPMC coeffiecients the overlap is between the known positives' minimum value and the known negatives' maximum value. Emulating previous trends, the narrowest ranges for all three models belong to the training set. For the known positives, EASI WO had the narrowest range (0.949-0.9999), followed closely by EASI (0.911-0.9999) and lastly the consensus model with a much wider range (0.599-0.9991).

Both EASI and consensus models had overlapping ranges between the known positives and pseudoallococaine and pseudococaine. Additionally, EASI and consensus models also have

problems differentiating between cocaine and allococaine. Lastly, the consensus model presented overlapping ranges between the known positives and ecgonine methyl ester.

For EASI, **Figure 27** makes it clear that there is no overlap between the known positives and most of the compounds. This means that it is possible to set an errorless threshold to separate cocaine from ecgonine methyl ester, fentanyl, heroin, hydromorphone and methamphetamine. However, **Figure 29** shows that complete overlap between cocaine and its diastereomers makes it impossible to select a threshold that results in errorless classification.

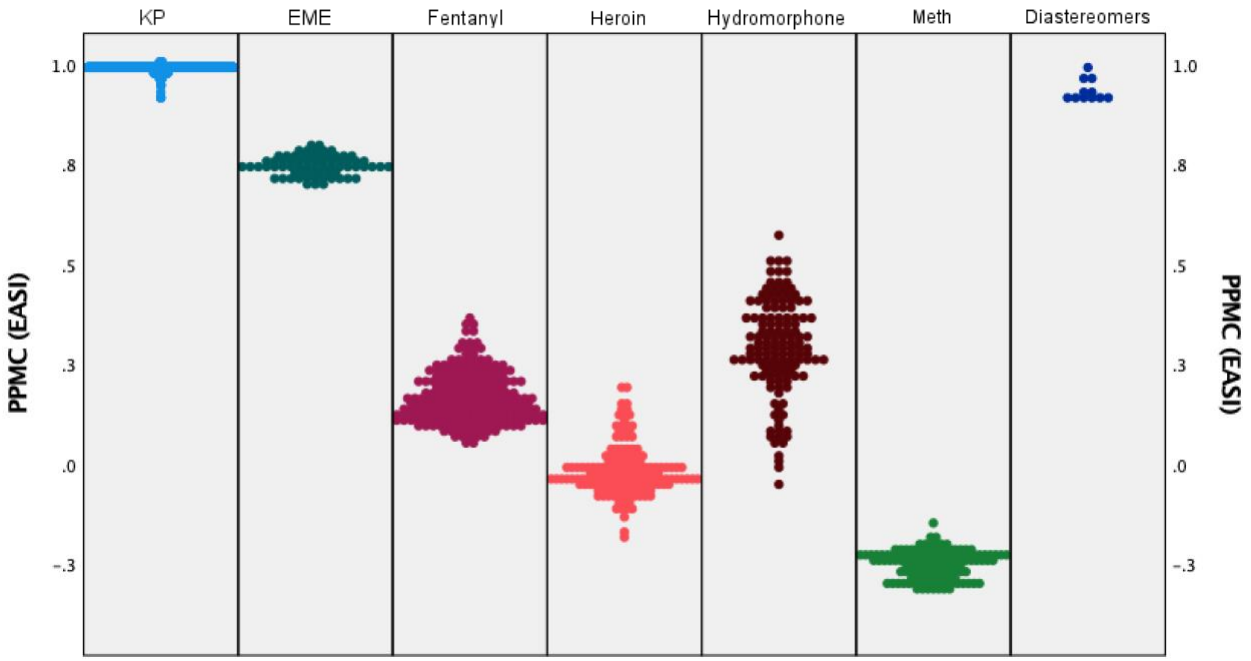

Figure 27. Population pyramid from EASI using the PPMC values. N_{KP}=1478, N_{EME}=69, $N_{fentanyl}$=216, N_{heroin}=158, $N_{hydromorphone}$=134, N_{meth}=133, $N_{diastereomers}$=11.

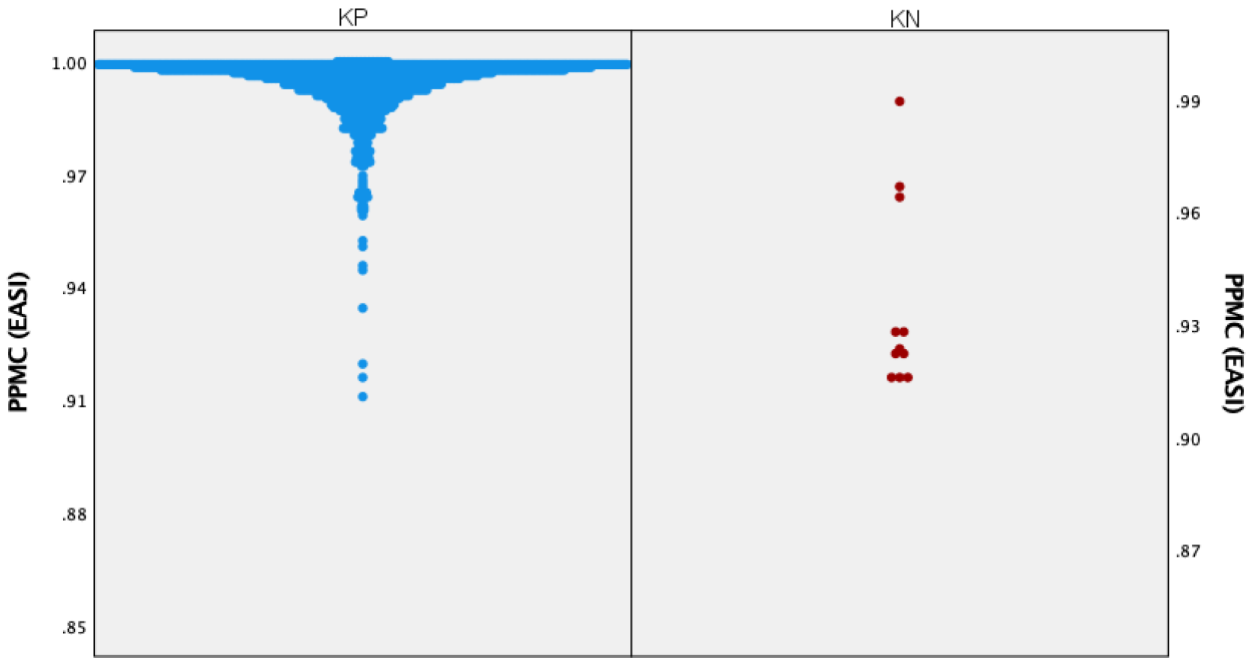

Figure 28. Close-up view of the population pyramid from EASI using the PPMC values. The left distribution (in blue) shows the known positives (N=1478), and the right distribution (in red) represents the known negatives (N=721).

For EASI WO, the different model and the penalization process in general decreased the PPMC values of all the known negatives and caused some of these distributions to divide into more than one cluster, which facilitates determining an errorless threshold for a binary classifier (**Figure 29**). The distribution of known positives using EASI WO was also tighter in comparison to EASI because of the penalization (**Figure 30**), which helped distinguish cocaine from allococaine. Similar to EASI, there is still complete overlap with pseudococaine and pseudoallococaine that hinders the selection of an errorless binary classifier.

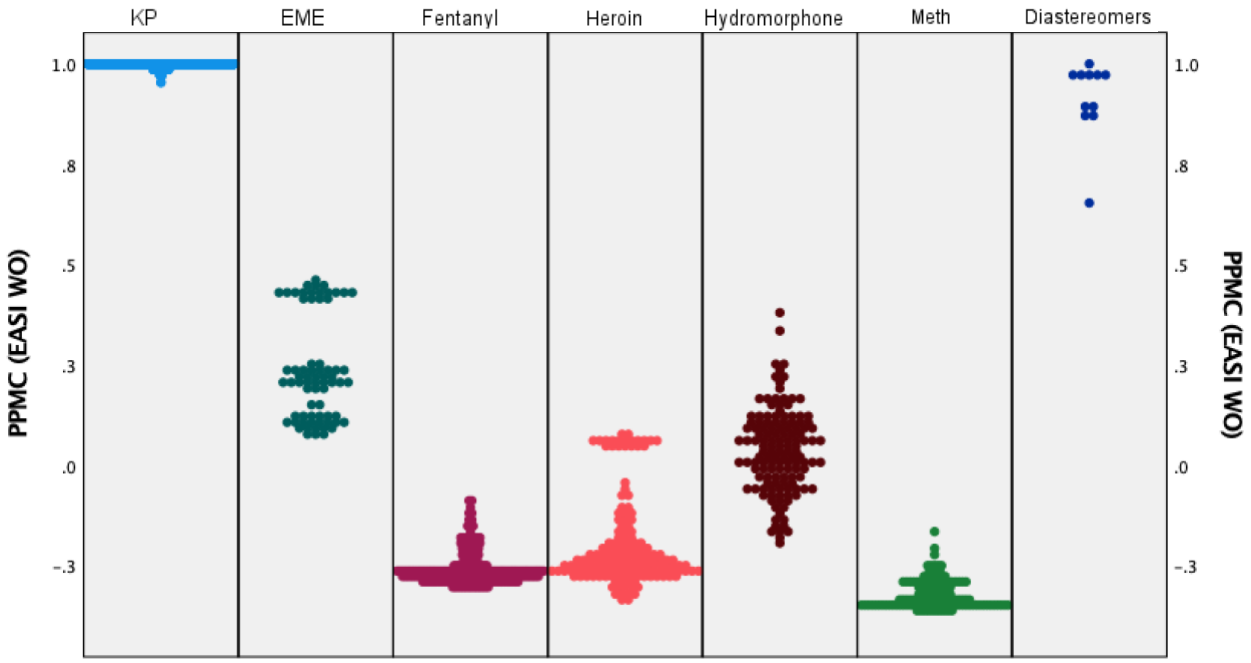

Figure 29. Population pyramid from EASI WO using the PPMC values. $N_{KP=}1478$, $N_{EME}=69$, $N_{fentanyl}=216$, $N_{heroin}=158$, $N_{hydromorphone}=134$, $N_{meth}=133$, $N_{diastereomers}=11$.

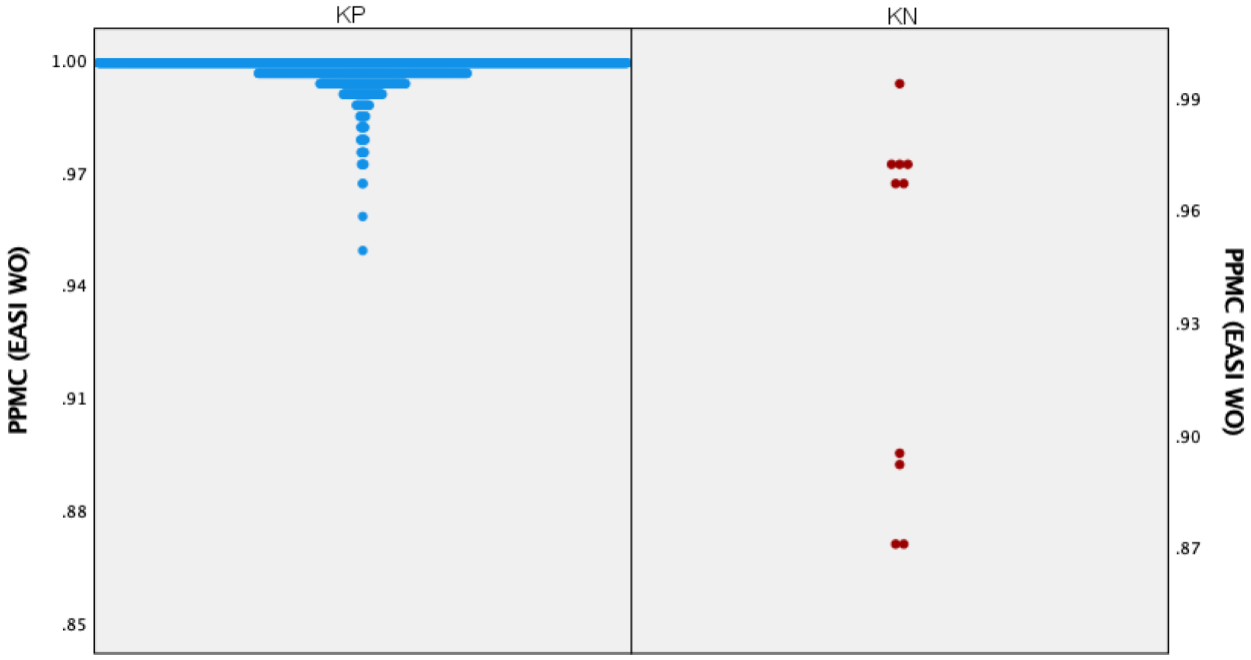

Figure 30. Close-up view of the population pyramids from EASI WO using the PPMC values. The left distribution (in blue) shows the known positives (N=1478), and the right distribution (in red) represents the known negatives (N=721).

For the consensus model, the distribution of known positives is much wider than for either EASI or EASI WO, which results in complete overlap with ecgonine methyl ester and the cocaine diastereomers. This means that there is no threshold that can prevent at least one false negative for cocaine relative to its diastereomers or ecgonine methyl ester. However, it is possible to select a threshold that prevents false positives or false negatives when analyzing cocaine against the known negatives of fentanyl, heroin, hydromorphone and methamphetamine. (**Figure 31** and **Figure 32**).

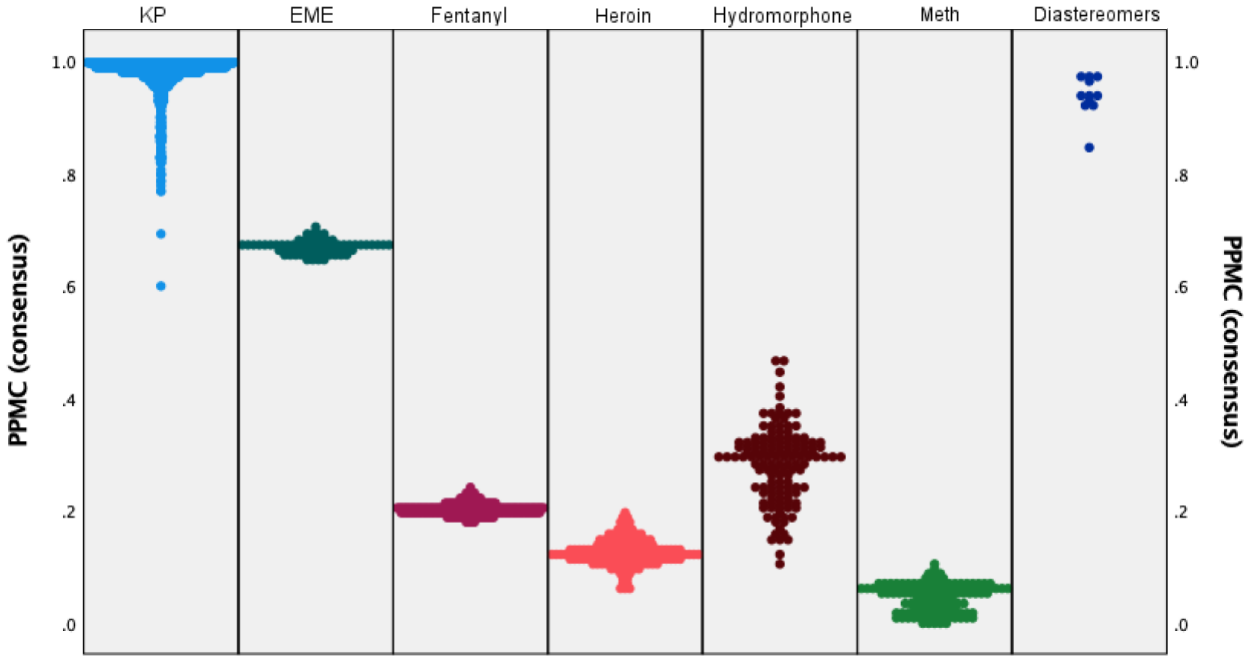

Figure 31. Population pyramid from the consensus model using the PPMC values. N_{KP}=1478, N_{EME}=69, $N_{fentanyl}$=216, N_{heroin}=158, $N_{hydromorphone}$=134, N_{meth}=133, $N_{diastereomers}$=11.

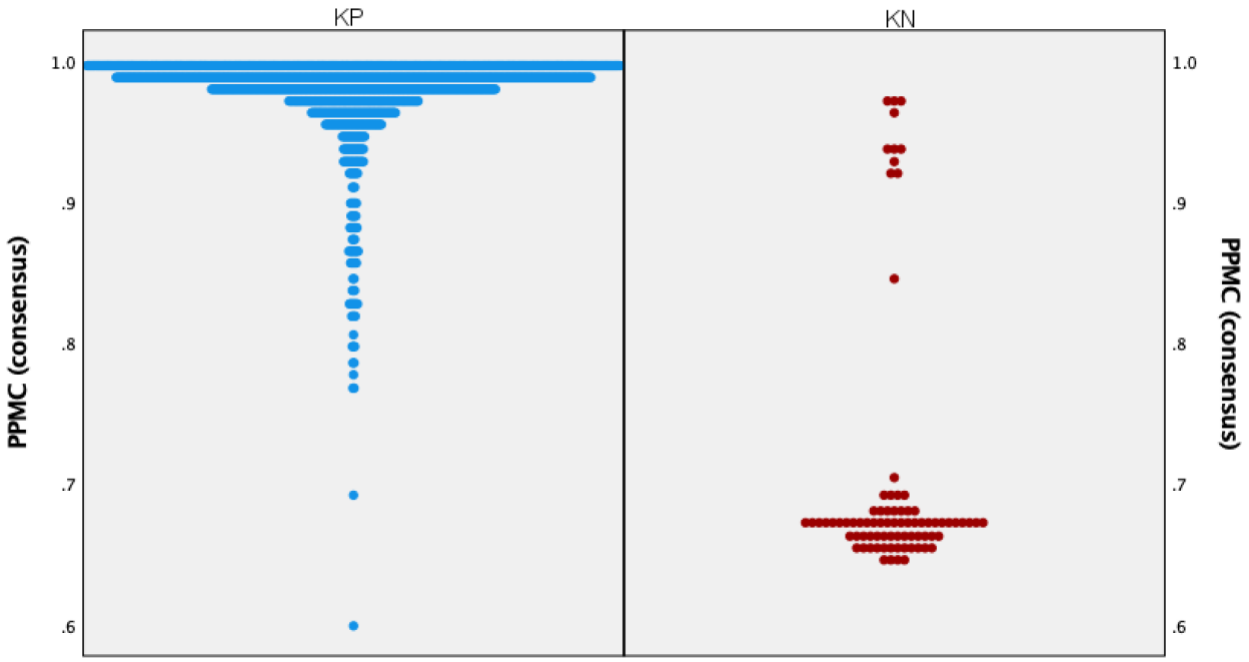

Figure 32. Close-up view of the population pyramid from the consensus model using the PPMC values. The left distribution (in blue) shows the known positives (N=1478), and the right distribution (in red) represents the known negatives (N=721).

To compare binary classifiers that use the PPMC values derived from EASI, EASI WO and consensus approach, **Figure 33** shows ROC curves resulting from all cocaine spectra as the known positives and all other spectra as known negatives, including the cocaine diastereomers.

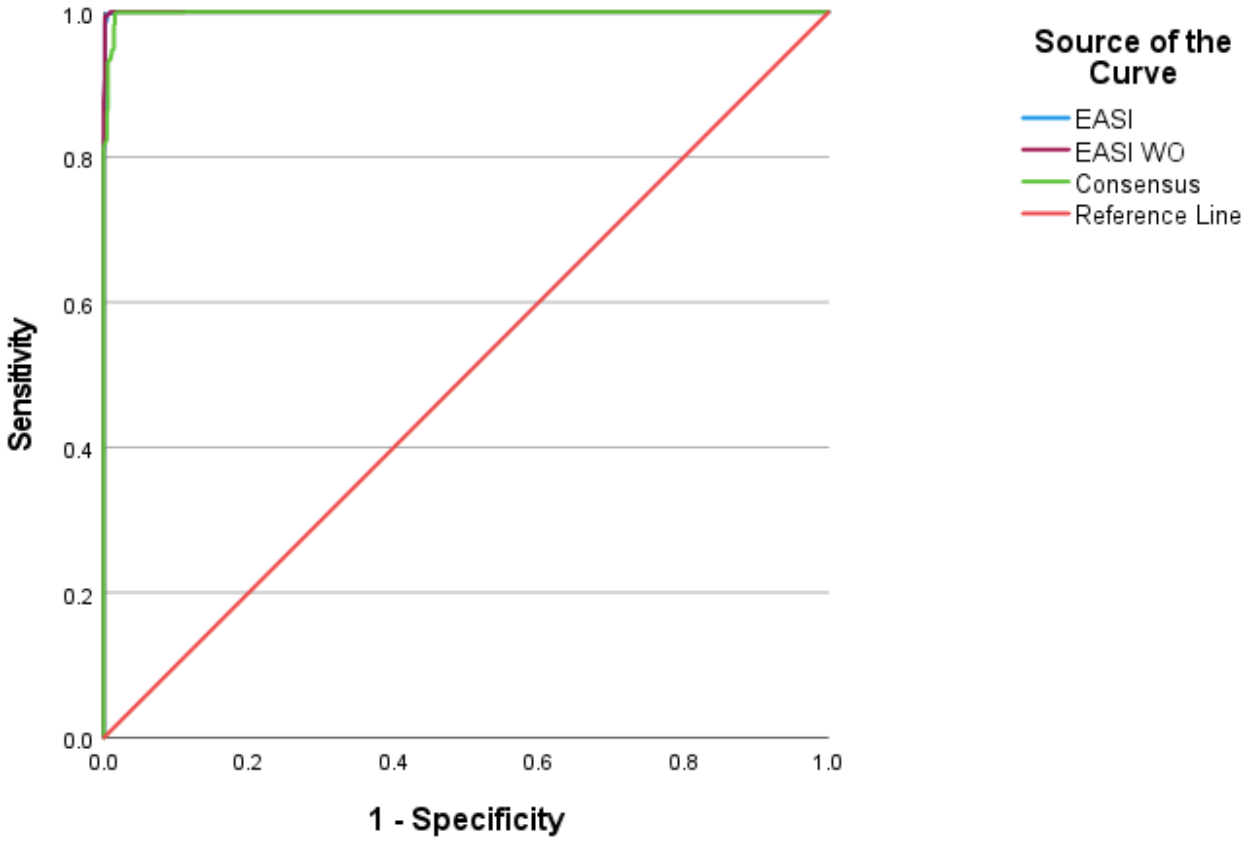

Figure 33. ROC curve generated using the PPMC values of all known positives (N=1478) and negatives (N=721).

Table 21 shows the area under the curve of a receiving-operator characteristic (ROC) using PPMC values as the classification metric. EASI and EASI WO provided almost indistinguishable AUCs of 0.9997 and 0.9998, respectively, and the consensus approach provided an AUC of 0.9985, which was similar to the results using MARs and Euclidean distances. As before, all three models are adequate to distinguish cocaine from most of the known positives.

Table 21. The AUCs generated using the PPMC values and all known negatives and positives.

Model	AUC
EASI	0.9997
EASI WO	0.9998
Consensus	0.9985

To further assess the performance of the models, we also generated ROC curves using only the cocaine diastereomers as the known negatives (**Figure 34**).

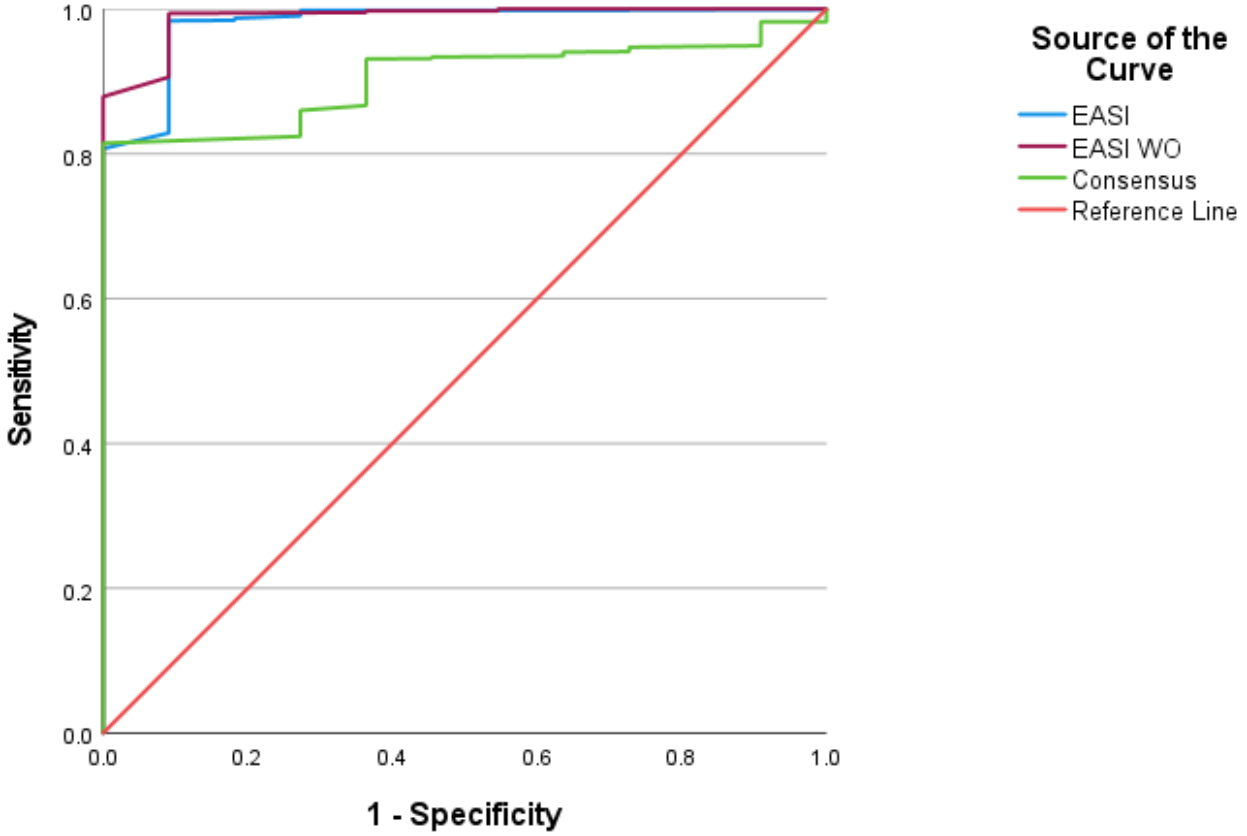

Figure 34. ROC curve generated using the PPMC values of all known positives (N=1478) and, pseudococaine, allococaine and pseudoallococaine as known negatives (N=11).

In this case, the models have more significant differences. **Table 22** shows that EASI WO still has the highest AUC with a 98.8% probability that the classifier will rank a randomly selected cocaine sample higher than a randomly selected known negative. The EASI is not far behind with an AUC of 98.0%, whereas the consensus model only has a probability of 90.4%.

Table 22. The AUCs generated using the PPMC values and all known positives and the cocaine diastereomers as the known negatives.

Model	AUC
EASI	0.9798
EASI WO	0.9883
Consensus	0.9037

4.2.5 NIST scores calculations and graphs

We calculated the NIST score as the weighted dot products using the optimized values as described by Stein et. al.[16] The results are seen in **Table 23**.

Table 23. Minimum and maximum NIST scores by compound and model. Bold values are the smallest NIST scores for known positives. Underlined values are the largest NIST scores for known negatives that overlap with the distribution of known positives.

	EASI		**EASI WO**		**Consensus**	
	Min	Max	Min	Max	Min	Max
KPs (training set)	977.8	999.0	966.9	999.0	968.2	999.0
KPs (test set)	**760.8**	999.0	**766.3**	999.0	**909.8**	998.9
Pseudoallococaine	982.9	987.7	979.9	987.6	990.5	994.4
Pseudococaine	982.9	996.6	978.0	995.9	984.4	995.6
Allococaine	989.5	989.5	991.1	991.1	991.7	991.7
Ecgonine methyl ester	798.9	925.1	155.7	350.8	386.7	498.5
Fentanyl	108.3	424.5	5.4	301.2	102.0	421.1
Heroin	215.7	579.3	116.6	460.6	407.9	530.4
Hydromorphone	21.5	547.1	86.5	575.0	191.1	509.9
Methamphetamine	40.6	77.2	4.8	17.8	63.3	90.3

In this case, the narrowest spread for the training set belonged to EASI (977.8-999.0), followed by the consensus model (968.2-999.0), and EASI WO (966.9-999.0). However, the spread for the test set did not adhere to the same trend. Instead, the consensus model had the narrowest spread (909.8-998.9), followed by EASI WO (766.3-999.0), and EASI (760.8-999.0). These scores indicate that by giving different weights to different *m/z* values and abundances, the dot product can be weighted to favor the consensus approach, upon which the NIST scores were essentially optimized.

Both EASI WO and the consensus model had no problems distinguishing cocaine from ecgonine methyl ester, fentanyl, heroin, hydromorphone, and methamphetamine based on the non-overlapping ranges of their NIST scores. However, all three models have complete overlapping ranges between cocaine and its diastereomers. In addition to the cocaine diastereomers, EASI also

presented a complete overlap between the NIST score ranges of cocaine and ecgonine methyl ester (**Figure 35**).

Figure 35 shows the population distribution for the NIST scores using EASI. It shows that although the bulk of the data for the cocaine is clustered around 999, there is a narrow tail or cocaine spectra with poorer NIST scores spreading from 999 to 760. This tail results in the complete overlap of cocaine and ecgonine methyl ester and eliminates the possibility of an errorless classifier. Additionally, in **Figure 36** the end of the cocaine tail consists of a few datapoints. These datapoints represent cocaine spectra that have *m/z* 182 as their base peak, which causes the NIST score to decrease.

Figure 36 shows that, unlike the distribution of ecgonine methyl ester, the distribution of cocaine diastereomers overlaps with a significant proportion of the cocaine distribution. This means that there is no threshold that would result in an errorless classification.

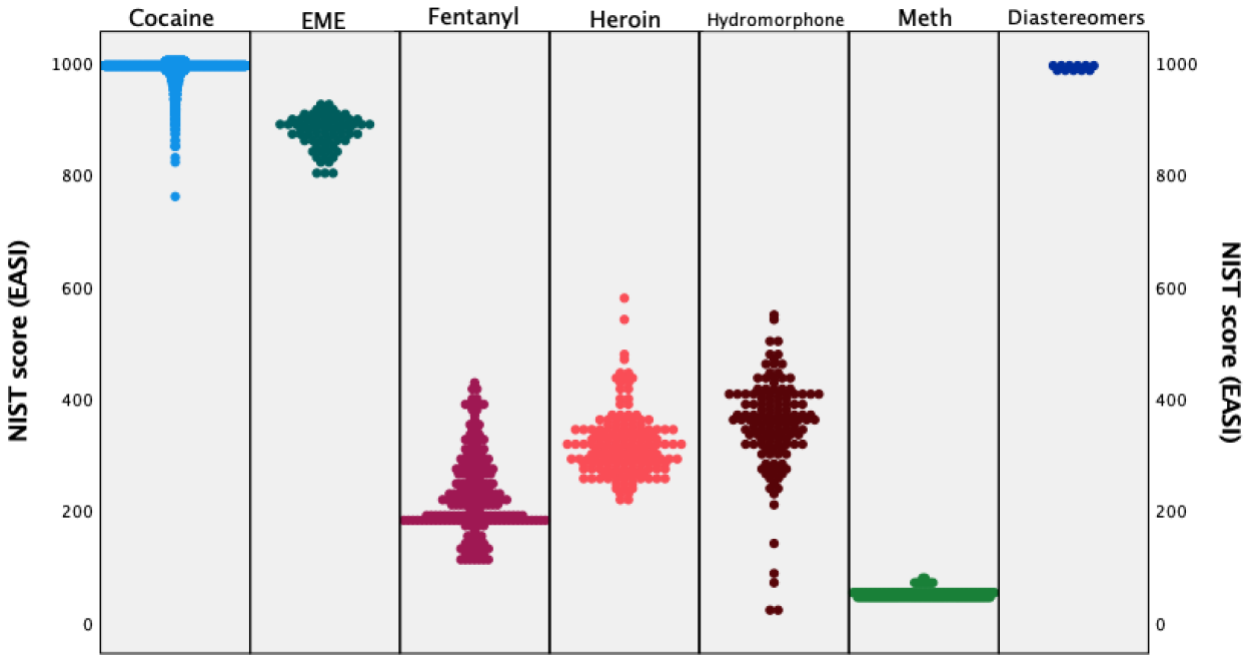

Figure 35. Population pyramid from EASI using the NIST scores. N_{KP}=1478, N_{EME}=69, $N_{fentanyl}$=216, N_{heroin}=158, $N_{hydromorphone}$=134, N_{meth}=133, $N_{diastereomers}$=11.

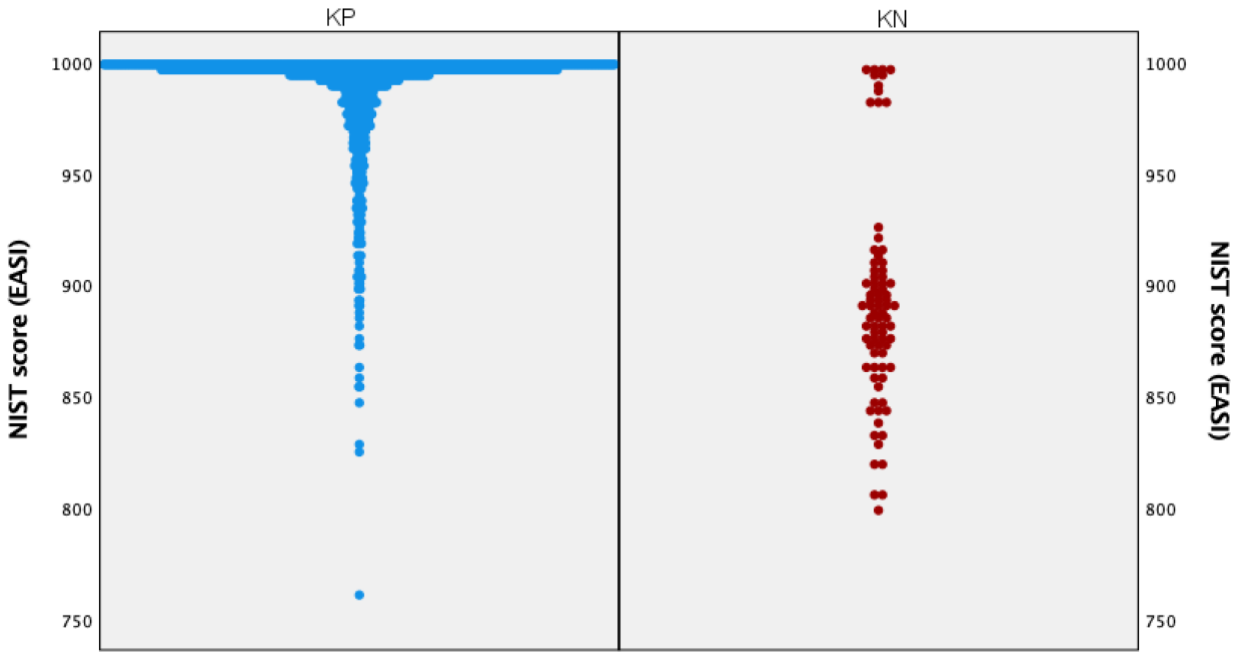

Figure 36. Close-up view of the population pyramid from EASI using the NIST scores. The left distribution (in blue) shows the known positives (N=1478), and the right distribution (in red) represents the known negatives (N=721).

Similar to EASI, the cocaine distribution using EASI WO also had a tail, but unlike EASI, this tail does not include the distribution of ecgonine methyl ester scores (**Figure 37**). Thanks to the penalization process, the NIST scores for the ecgonine methyl ester went from 798.9-925.1 with EASI to 155.7-350.8 with EASI WO. For EASI WO, the only overlap is therefore between cocaine and its diastereomers. **Figure 38** shows the complete overlap between these compounds, which demonstrates the inability to set an errorless threshold.

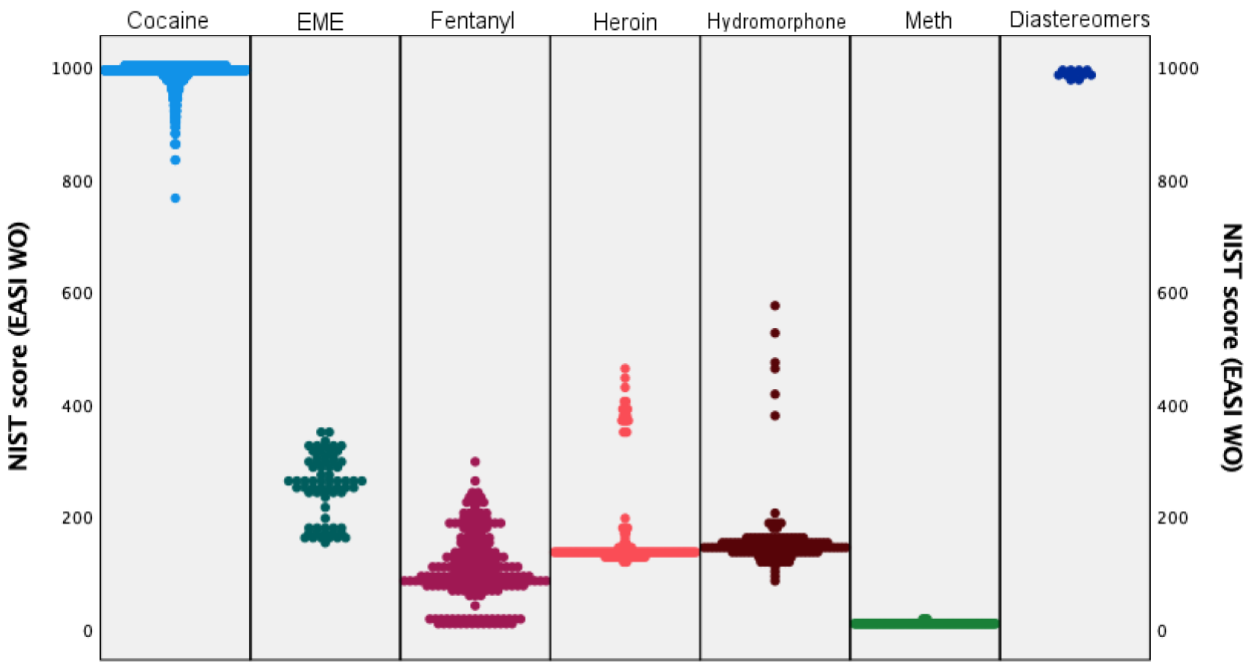

Figure 37. Population pyramid from EASI WO using the NIST scores. $N_{KP=}1478$, $N_{EME}=69$, $N_{fentanyl}=216$, $N_{heroin}=158$, $N_{hydromorphone}=134$, $N_{meth}=133$, $N_{diastereomers}=11$.

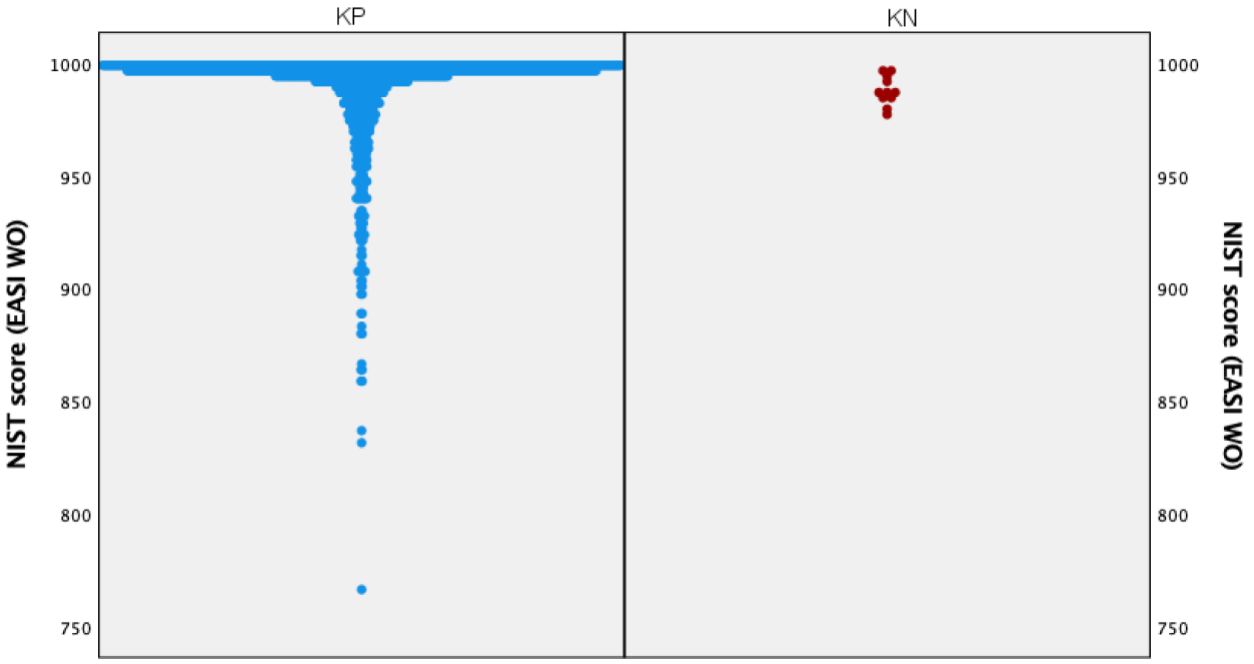

Figure 38. Close-up view of the population pyramid from EASI WO using the NIST scores. The left distribution (in blue) shows the known positives (N=1478), and the right distribution (in red) represents the known negatives (N=721).

As stated before, the consensus model coupled with the NIST scoring was able to clearly separate cocaine from ecgonine methyl ester, fentanyl, heroin, hydromorphone, and methamphetamine. But, like EASI and EASI WO, the consensus approach had overlapping ranges between cocaine and its diastereomers (**Figure 39**). Although the cocaine distribution has a tail (like EASI and EASI WO), the resulting NIST scores formed a tighter cluster around 999 than with either EASI or EASI WO. This means that it can be easier to set a threshold that minimizes the false positives.

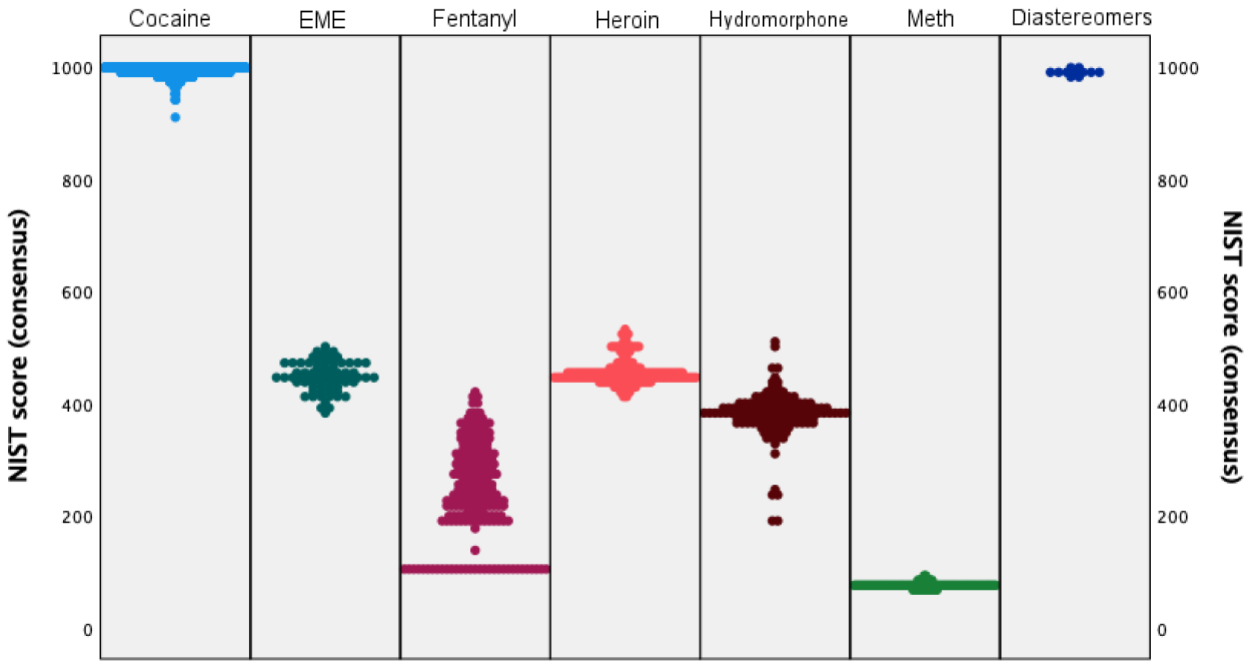

Figure 39. Population pyramid from the consensus model using the NIST scores. N_{KP}=1478, N_{EME}=69, $N_{fentanyl}$=216, N_{heroin}=158, $N_{hydromorphone}$=134, N_{meth}=133, $N_{diastereomers}$=11.

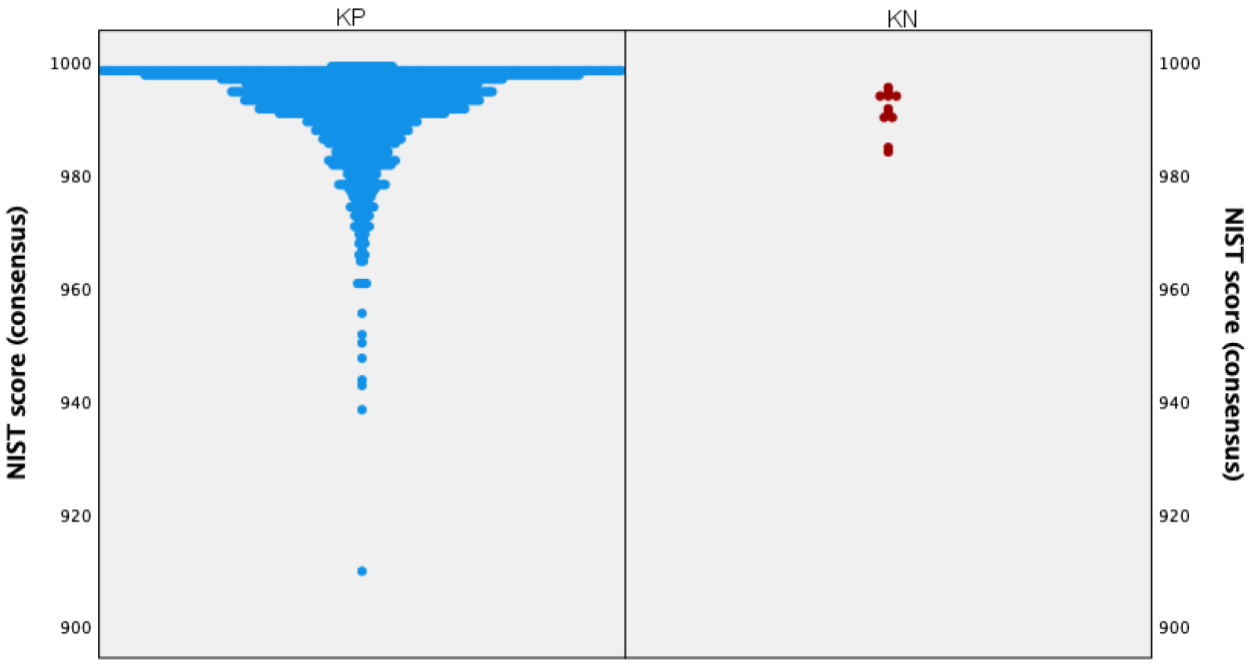

Figure 40. Close-up view of the population pyramid from the consensus model using the NIST scores. The left distribution (in blue) shows the known positives (N=1478), and the right distribution (in red) represents the known negatives (N=721).

The ROC curves generated using the NIST scores for all known positives and all known negatives are shown in **Figure 41**. All three models had comparable AUCs (**Table 24**). The lowest AUC was 0.980 for EASI WO, followed by a tie between EASI and the consensus model with an AUC of 0.994. This shows that the three models have a high chance of correctly classifying cocaine against drugs that are not its diastereomers.

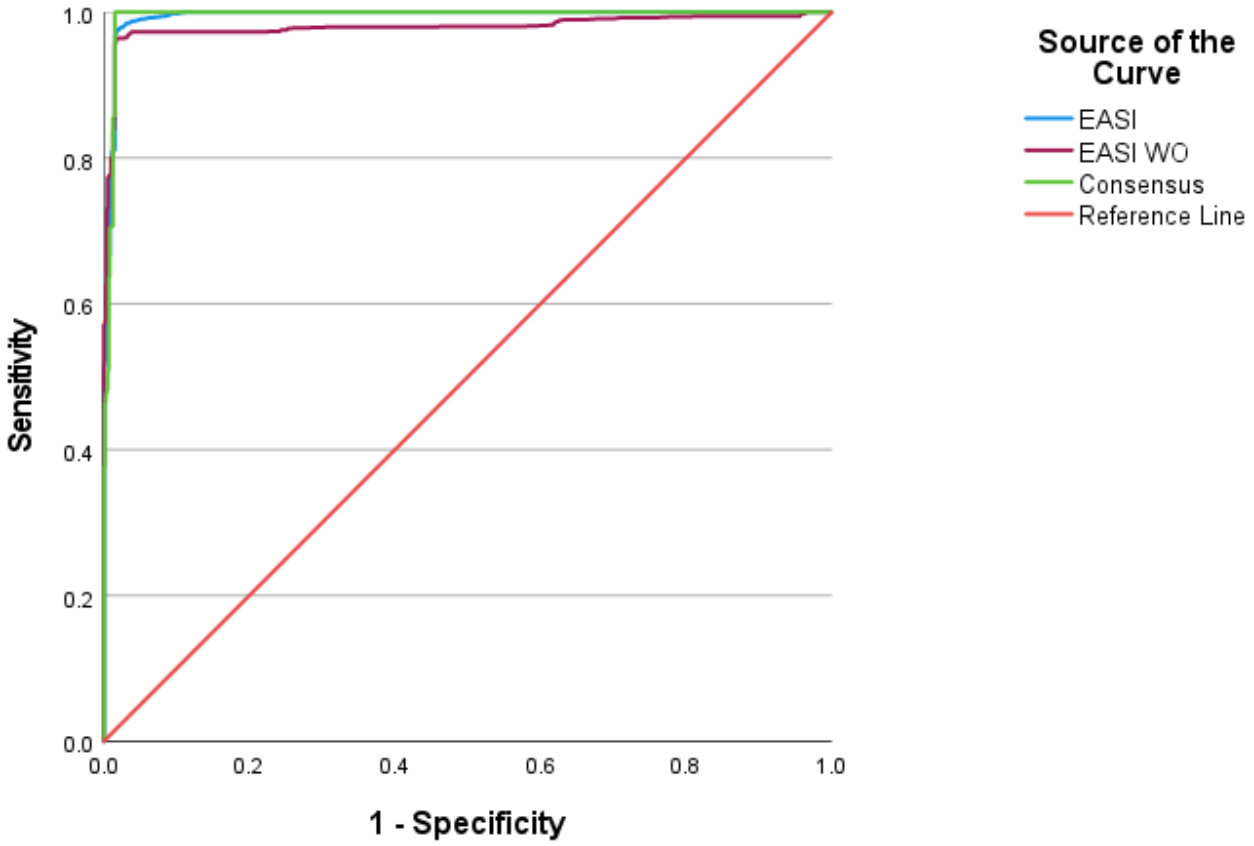

Figure 41. ROC curve generated using the NIST scores of all known positives (N=1478) and negatives (N=721).

Table 24. The AUCs generated using the NIST scores and all known negatives and positives.

Model	AUC
EASI	0.9941
EASI WO	0.9797
Consensus	0.9941

When analyzing the ROC curves generated using the NIST scores and the cocaine diastereomers as the known negatives (**Figure 42**) the performance is worse than when the easy-to-distinguish known negatives are removed (e.g., **Figure 41**). This deterioration is consistent with all the measures of spectral comparison and is more evident if we look at the AUCs in **Table 25**. The model with the highest AUC was EASI WO with 0.739, whereas the consensus model only has a chance of a correct identification of 61.6%. The EASI had an AUC of 0.676.

These results show that even though the consensus model had one of the highest AUCs using all known negatives, when trying to classify cocaine among its diastereomers it does not perform better than any of the EASI models. Note that for distinguishing cocaine from its diastereomers, NIST scores performed significantly worse than any of the other spectral comparison methods. This outcome demonstrates that the success in distinguishing cocaine from its diastereomers is not enhanced by favoring high mass ions or low-abundance ions. In fact, Casale et al. and Smith have shown that cocaine diastereomers are best resolved from cocaine using peak ratios at *m/z* 94:96 and *m/z* 152:150.[51,52]

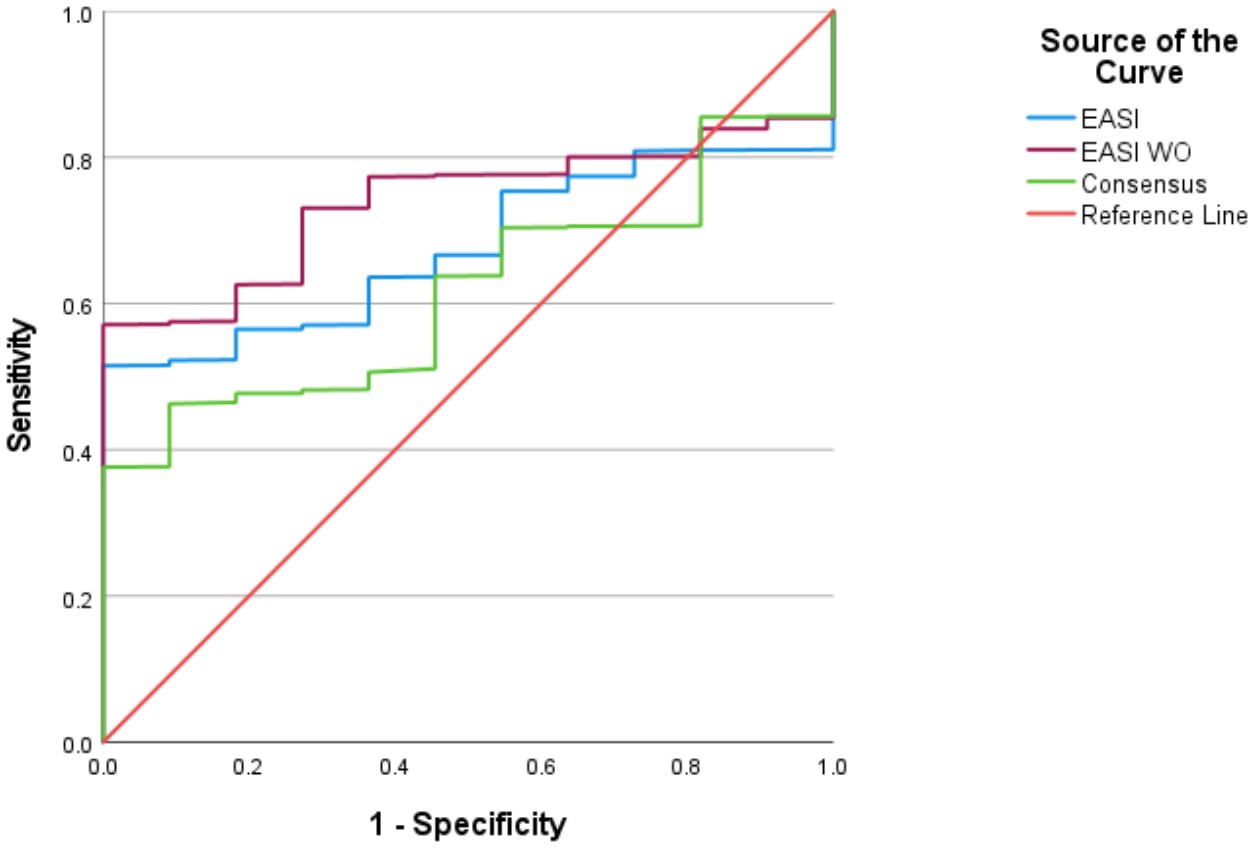

Figure 42. ROC curve generated using the NIST scores of all known positives (N=1478) and, pseudococaine, allococaine and pseudoallococaine as known negatives (N=11).

Table 25. The AUCs generated using the NIST scores and all known positives and the cocaine diastereomers as the known negatives.

Model	AUC
EASI	0.6758
EASI WO	0.7386
Consensus	0.6158

4.2.6 Mahalanobis distances calculations and graphs

As a final measure of spectral comparison, the Mahalanobis distances were calculated as described before without any linear modeling needed, using the measured abundances of all 20 *m/z* values (**Table 26**) in the training set as the basis for establishing the covariance matrix.

Table 26. Minimum and maximum Mahalanobis distances by compound. Bold values are the largest Mahalanobis distances for known positives. Underlined values are the smallest Mahalanobis distances for known negatives that overlap with the distribution of known positives.

	Min	Max
KPs (training set)	2.04	73.30
KPs (test set)	1.69	**261.68**
Pseudoallococaine	<u>123.40</u>	177.14
Pseudococaine	<u>95.28</u>	301.65
Allococaine	<u>182.89</u>	182.89
Ecgonine methyl ester	770.53	995.67
Fentanyl	323.49	641.18
Heroin	317.08	1118.40
Hydromorphone	456.52	1267.53
Methamphetamine	304.01	315.31

The Mahalanobis distances followed the basic trends discussed above for the other spectral comparison methods. The training set range had the smallest maximum distance, with a value of 73.30. The validation set had a larger maximum Mahalanobis distance of 261.68, which is high enough to include all three cocaine diastereomers and make it impossible to set an errorless threshold.

These ranges are better visualized in **Figure 43**. This figure shows that the distribution for Mahalanobis distances for the training set of cocaine also had tail, as a result of KPs in the test set with a different base peak than the training set. **Figure 44** shows that the presence of this tail causes overlap between cocaine and its diastereomers. This tail could also cause a problem when trying to distinguish between cocaine and methamphetamine, if the Mahalanobis distance start to overlap.

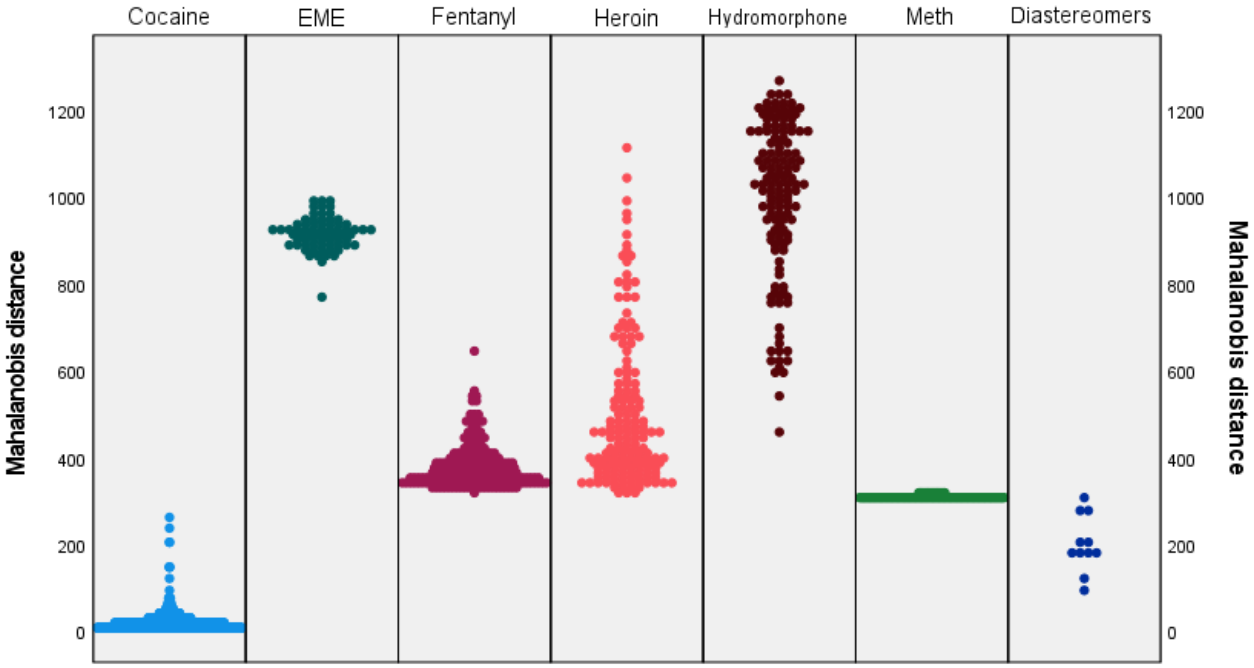

Figure 43. Population pyramid using the Mahalanobis distances. N_{KP}=1478, N_{EME}=69, $N_{fentanyl}$=216, N_{heroin}=158, $N_{hydromorphone}$=134, N_{meth}=133, $N_{diastereomers}$=11.

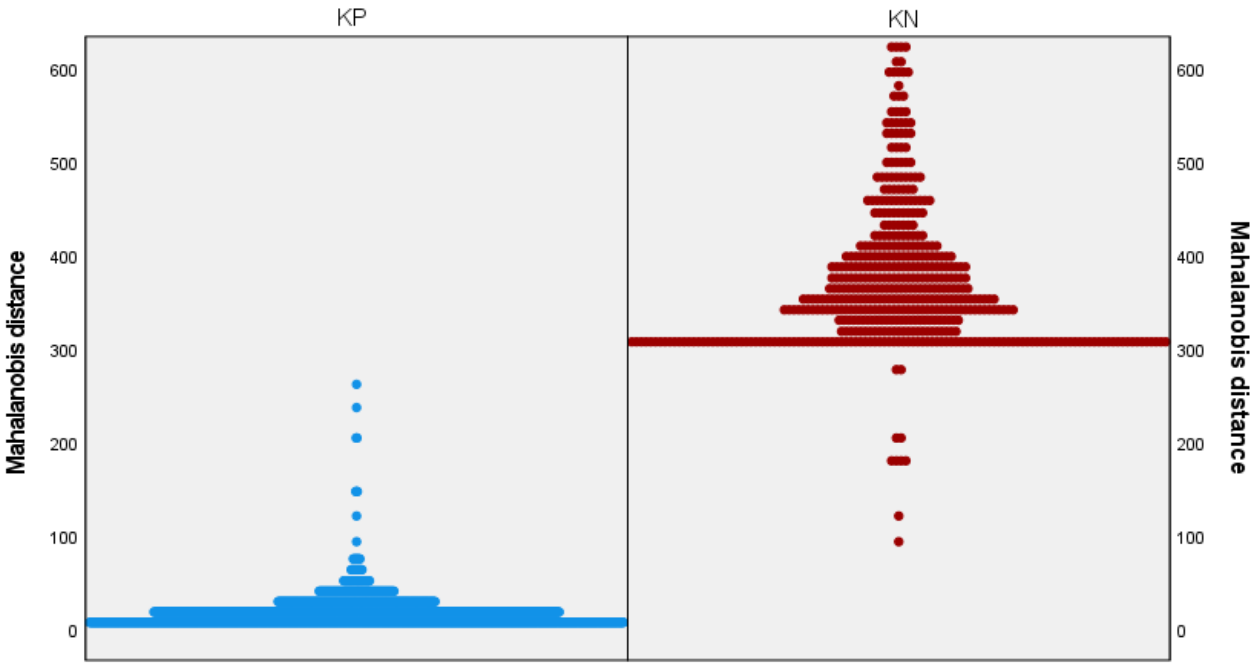

Figure 44. Close-up view of the population pyramid using the Mahalanobis distances. The left distribution (in blue) shows the known positives (N=1478), and the right distribution (in red) represents the known negatives (N=721).

The ROC curve generated using all known negatives and the Mahalanobis distances as the decision metric shown in **Figure 45** had an area under the curve of 0.999967. This AUC reflects the excellent ability of this classifier to easily distinguish most cocaine spectra from most known negatives.

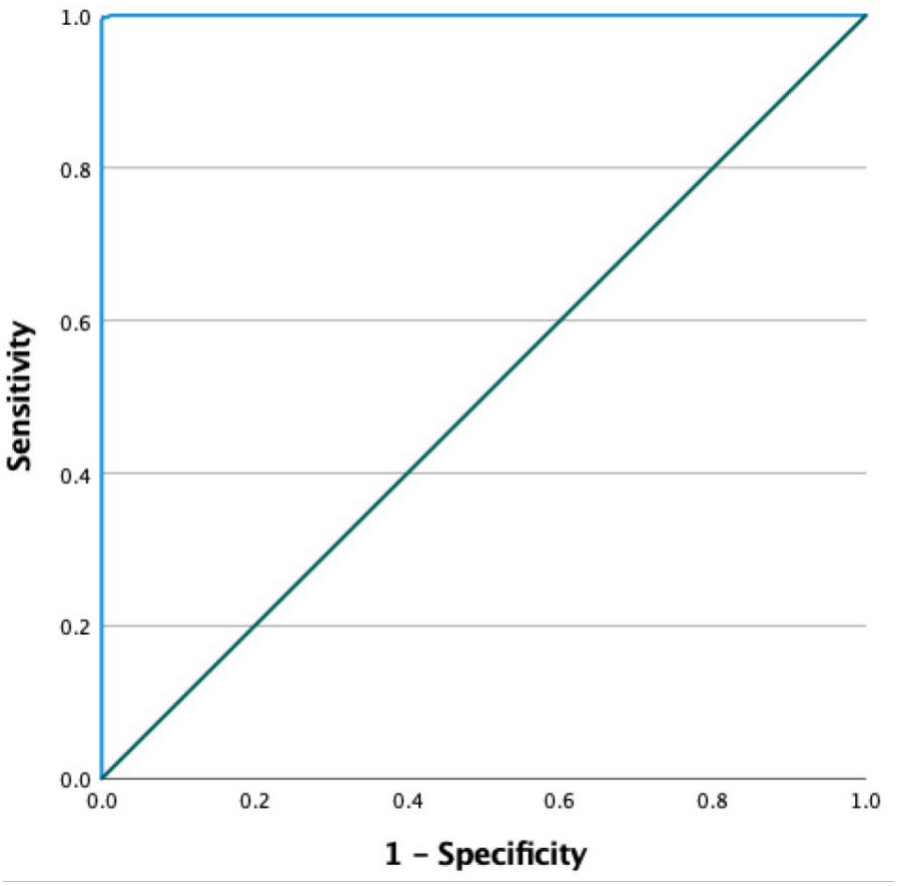

Figure 45. ROC curve generated using the Mahalanobis distance of all known positives (N=1478) and negatives (N=721). The area under the curve was 0.999967.

The ROC curve shown in **Figure 46** was generated using the cocaine diastereomers as the only known negatives and it resulted in an area under the curve of 0.997847, meaning that it has a 99.8% chance of correctly identifying cocaine from its diastereomers. The slight decrease in AUC from the previous ROC curve is logical considering the overlapping ranges between cocaine and its diastereomers.

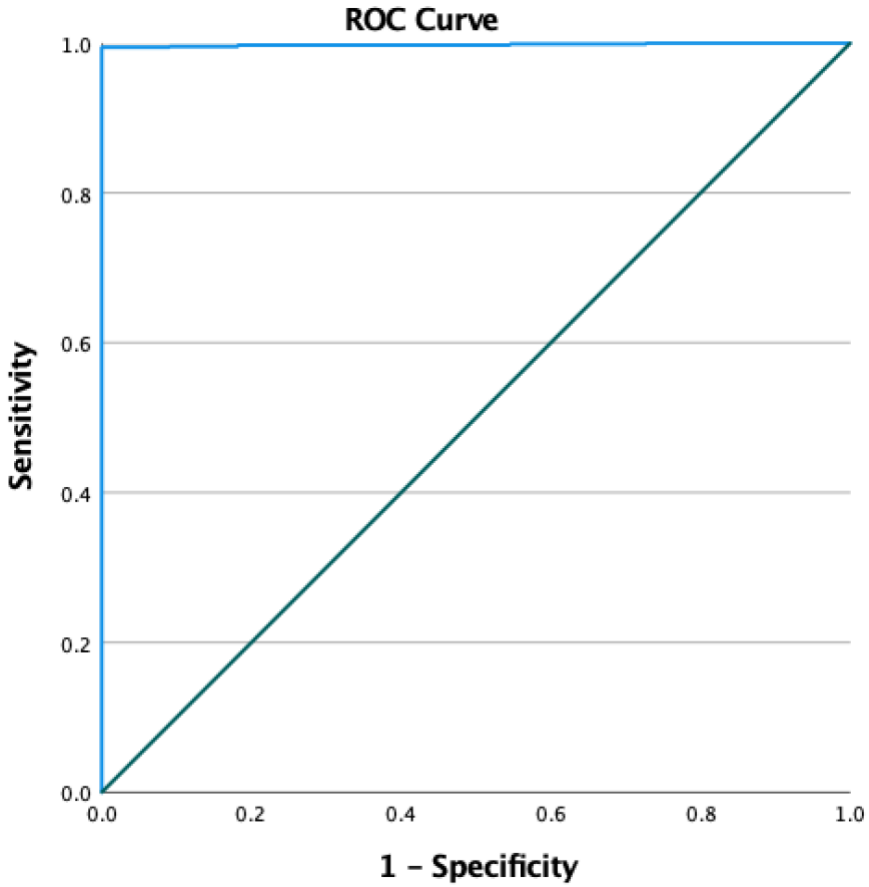

Figure 46. ROC curve generated using the Mahalanobis distance of all known positives (N=1478) and only pseudococaine, allococaine and pseudoallococaine as known negatives (N=11). The area under the curve was 0.997847.

4.2.7 Model and metrics comparison

To assess and compare the three models coupled to the five metrics discussed above, **Table 27** shows the true positives, true negatives, false positives, false negatives, true positive rate, false positive rate and accuracy at a threshold that results in a 0% false positive rate (or 100% true negative rate). This 0% false positive rate was chosen given that the GC-MS technique is a confirmatory test, and it is important to minimize the false positive rate in these tests.

Table 27 shows that within each metric of spectral comparison, one of the GLM-based algorithms (EASI or EASI WO) results in greater accuracy and true positive rate than the

consensus approach. The Mahalanobis distance as a binary classifier provides the highest accuracy and true positive rate overall with values of 99.6% and 99.5%, respectively. However, if the Mahalanobis test was used in conjunction with a Hotelling's T-squared distribution, many of the cocaine spectra in the training and test set would be considered outliers and would be classified as false negatives. The GLM-based models in EASI have the added advantage of extrapolation outside of the training set's variance. The lowest accuracies for each model were obtained using the NIST scores, which can be partly attributed to optimization of this metric to the consensus approach. Excluding the NIST scores results, more than 85% of the known positives can be correctly identified as cocaine with zero false positives using either of the GLM-based algorithms.

After the Mahalanobis distance, the best classification rates were obtained using EASI coupled with the mean absolute residuals as a binary classifier. A threshold of 2.60 resulted in zero false positives and 49 false negatives, for an accuracy of 97.8%. These results align closely with EASI models based on a completely different training set and a test set from a third laboratory.[43]

Table 27. Confusion matrix with all known negatives at 0% FPR.

	Mean absolute residual			**PPMC**			**Euclidean distance**			**NIST score**			**Mahala nobis**
	EASI	**EASI WO**	**CNS**	**EASI**	**EASI WO**	**CNS**	**EASI**	**EASI WO**	**CNS**	**EASI**	**EASI WO**	**CNS**	
Threshold	2.60	2.49	4.11	0.990	0.993	0.969	14.41	14.17	28.01	996.6	995.9	995.6	95.28
TPs	1429	1399	1255	1222	1314	1208	1391	1334	1156	762	845	556	1470
TNs	721	721	721	721	721	721	721	721	721	721	721	721	721
FPs	0	0	0	0	0	0	0	0	0	0	0	0	0
FNs	49	79	223	256	164	270	87	144	322	716	633	922	8
TPR	96.7	94.7	84.9	82.7	88.9	81.7	94.1	90.3	78.2	51.6	57.2	37.6	99.5
FNR	3.3	5.35	15.09	17.32	11.10	18.27	5.89	9.74	21.79	48.44	42.83	62.38	0.54
Accuracy	97.8	96.4	89.9	88.4	92.5	87.7	96.0	93.5	85.4	67.4	71.2	58.1	99.6

Using only the diastereomers as known negatives while maintaining a 0% false positive rate results in decreased accuracies and increased false negative rates for all models (**Table 28**). However, the GLM-based models still outperform the consensus approach using any of the

spectral comparison metrics. The EASI coupled with the mean absolute residuals still has the highest accuracy (96.7%) after the Mahalanobis approach (99.5%), followed by EASI WO coupled with the mean absolute residuals (94.7%).

Table 28. Confusion matrix with the cocaine diastereomers as known negatives at 0% FPR.

	Mean absolute residual			**PPMC**			**Euclidean distance**			**NIST score**			**Mahala nobis**
	EASI	**EASI WO**	**CNS**	**EASI**	**EASI WO**	**CNS**	**EASI**	**EASI WO**	**CNS**	**EASI**	**EASI WO**	**CNS**	
Threshold	2.60	2.49	4.11	0.990	0.993	0.969	14.41	14.17	28.01	996.6	995.9	995.6	95.28
TPs	1429	1399	1255	1222	1314	1208	1391	1334	1156	762	845	556	1470
TNs	11	11	11	11	11	11	11	11	11	11	11	11	11
FPs	0	0	0	0	0	0	0	0	0	0	0	0	0
FNs	49	79	223	256	164	270	87	144	322	716	633	922	8
TPR	96.7	94.7	84.9	82.7	88.9	81.7	94.1	90.3	78.2	51.6	57.2	37.6	99.5
FNR	3.3	5.35	15.09	17.32	11.10	18.27	5.89	9.74	21.79	48.44	42.83	62.38	0.54
Accuracy	96.7	94.7	85.0	82.8	89.0	81.9	94.2	90.3	78.4	51.9	57.5	38.1	99.5

These results demonstrate the superiority of the GLM-based models (EASI and EASI WO) over the consensus approach. Additionally, they show the potential of the Mahalanobis distances as a binary classifier, albeit without the advantage of the ability to understand the behavior of individual ions or extrapolate models for specific *m/z* values from one set of data to another. The NIST scores represented the lowest accuracies with 51.9%, 57.5% and 38.1% for EASI, EASI WO and consensus approach, respectively. These low values can be due to the added artificial variance in the training set, which serves as a basis for the three models. Additionally, the NIST scores have been shown to be highly dependent on the optimization variables.[16]

The results of this project serve as a proof of concept for GLM-based model to predict ion abundances more accurately than the current consensus model. Additionally, EASI and EASI WO can be coupled with at least four different metrics to create binary classifiers that perform better than the consensus model. This provides freedom for users to select the metric that best suits their needs.

5. Conclusions and future work

When assessing the impact of the repeller voltage, focus lens voltage and the electron ionization energy, uncharacteristically large changes in these three factors only explained around 3% of the variance observed in the training set comprised of 389 spectra from a full-factorial design of experiments. Instead, more than 90% of the variance in the product ion abundances of an additional 28 spectra collected over ~2.5 minutes could be explained by random drift in the high vacuum. Similar results were shown by a previous lab member with the calibration gas PFTBA as the analyte. This study showed that the pressure caused significant changes in the branching ratios of cocaine, with some ions correlating more strongly with one another than others. These findings support the recent proposal that empirical correlations and anticorrelations between normalized ion abundances in replicate mass spectra are predicted by RRKM theory. The underlying statistical foundations provide a robust platform for empirical modeling using general linear models.

Each of the 20 most abundant ions in the training set of 389 cocaine spectra were iteratively conserved to be the dependent variable. Stepwise addition in SPSS was used to build general linear models for each *m/z* as the dependent variable with as many covariates as were necessary to explain the maximum about of variance in the dependent variable, without overfitting. The statistical validity of the GLM models were assessed through analysis of the residuals between modeled and measured abundances. GLM models were built with and without constants in the linear models, and with and without the inclusion of a penalty of 50% residual error each time an expected ion was missing in a spectrum.

EASI accounted for more than 80% of the variance in replicate spectra. Four different measures of similarity and dissimilarity were coupled to the GLM-based algorithms, EASI and EASI WO had consistently fewer false negatives when a 0% false positive threshold was set. The

EASI and EASI WO also outperformed the consensus approach when classifying cocaine based on ROC curves and AUCs, with both GLM-based algorithms generally presenting higher AUCs. EASI WO had slightly higher AUCs than EASI, but both GLM-based models resulted in reliably higher AUCs than the consensus approach. Additionally, Mahalanobis distances from the training set of 389 cocaine spectra was a very reliable binary classifier since this method produced the lowest false negative rate with zero false negatives.

Unlike Mahalanobis distance, the GLM-based models can be used to understand and extrapolate the measured data in one lab to the measured data in a second lab. GLM-based models can therefore account for variance in normalized mass spectra that cannot be controlled by factors such as drift in the high vacuum. This project serves as a proof-of-concept for GLM-based algorithms, showing their robustness and versatility. We were able to postulate different metrics that can be coupled to EASI or EASI WO to serve as binary classifiers, providing measures of accuracy for each one to inform the user.

The next step to continue this work is to expand the training set database to include cocaine spectra with *m/z* 182 as the base peak. This could potentially limit the outliers among the known positives in the test set and thus reduce the overlap between known positives and known negatives. In addition to this measure, it would also be beneficial to include more cocaine diastereomers in the database of known negatives because it is difficult to draw conclusions from only 11 spectra. To further try and differentiate the cocaine diastereomers, more importance could be given to *m/z* ratios 94:96 and 152:150, which have been previously shown to distinguish cocaine from its diastereomers.[52] Emphasizing these *m/z* values could potentially increase the accuracy of EASI and EASI WO.

6. References

(1) McLafferty, F. W.; Hertel, R. H.; Villwock, R. D. Probability Based Matching of Mass Spectra. *Org. Mass Spectrom.* **1974**, *9* (7), 690–702.

(2) Xie, C.; Yu, J. C.; Huang, S.; Gao, W.; Tang, K. A Novel Approach of Matching Mass-to-Charge Ratio for Compound Identification in Gas Chromatography–Mass Spectrometry. *J. AOAC Int.* **2019**, *102* (2), 638–645. https://doi.org/10.5740/jaoacint.18-0261.

(3) Stauffer, D. B.; McLafferty, F. W.; Ellis, R. D.; Peterson, D. W. Adding Forward Searching Capabilities to a Reverse Search Algorithm for Unknown Mass Spectra. *Anal. Chem.* **1985**, *57* (3), 771–773. https://doi.org/10.1021/ac00280a045.

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

(5) 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* (October 2018), 58–65. https://doi.org/10.1016/j.forc.2018.12.001.

(6) SWGDRUG. Recommendations for Code of Professional Practice, Education and Training, Methods of Analysis, and Quality Assurance. *Sci. Work. Gr. Anal. Seized Drugs* **2019**, No. 8, 83.

(7) Gross, J. H. *Mass Spectrometry*, Third Ed.; 2016; Vol. 1040.

(8) Paul, W. Electromagnetic Traps for Charged and Neutral p Articles. *Angew. Chemie Int. Ed. English* **1990**, *29* (7), 739–748. https://doi.org/10.1002/anie.199007391.

(9) McLafferty, F. W.; Gohlke, R. S. Spectral Data File Utilizing Machine Filing and Manual Searching. **1959**, *863*, 857–863.

(10) Crawford, L. R.; Morrison, J. D. Computer Methods in Analytical Mass Spectrometry Identification of an Unknown Compound in a Catalog. *Anal. Chem.* **1968**, *40* (10), 1464–1469. https://doi.org/10.1021/ac60266a027.

(11) Law, N. C.; Aandahl, V.; Fales, H. M.; Milne, G. W. A. Identification of Dangerous Drugs by Mass Spectrometry. *Clin. Chim. Acta* **1971**, *32* (2), 221–228. https://doi.org/10.1016/0009-8981(71)90336-6.

(12) Grotch, S. L. Matching of Mass Spectra When Peak Height Is Encoded to One Bit. *Anal. Chem.* **1970**, *42* (11), 1214–1222. https://doi.org/10.1021/ac60293a007.

(13) Knock, B. A.; Smith, I. C.; Wright, D. E.; Ridley, R. G.; Kelly, W. Compound Identification by Computer Matching of Low Resolution Mass Spectra. *Anal. Chem.* **1970**, *42* (13), 1516–1520. https://doi.org/10.1021/ac60295a035.

(14) Hertz, H. S.; Hites, R. A.; Biemann, K. Identification of Mass Spectra by Computer-Searching a File of Known Spectra. *Anal. Chem.* **1971**, *43* (6), 681–691. https://doi.org/10.1021/ac60301a009.

(15) Stein, S. E. Estimating Probabilities of Correct Identification from Results of Mass Spectral Library Searches. *J. Am. Soc. Mass Spectrom.* **1994**, *5* (4), 316–323. https://doi.org/10.1016/1044-0305(94)85022-4.

(16) 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.

(17) Moorthy, A. S.; Wallace, W. E.; Kearsley, A. J.; Tchekhovskoi, D. V.; Stein, S. E. Combining Fragment-Ion and Neutral-Loss Matching during Mass Spectral Library Searching: A New General Purpose Algorithm Applicable to Illicit Drug Identification.

Anal. Chem. **2017**, *89* (24), 13261–13268. https://doi.org/10.1021/acs.analchem.7b03320.
(18) Pesyna, G. M.; McLafferty, F. W.; Venkataraghavan, R.; Dayringer, H. E. Statistical Occurrence of Mass and Abundance Values in Mass Spectra. *Anal. Chem.* **1975**, *47* (7), 1161–1164. https://doi.org/10.1021/ac60357a050.
(19) Rasmussen, G. T.; Isenhour, T. L. The Evaluation of Mass Spectral Search Algorithms. **1979**, *19* (3).
(20) Mun, I. K.; Bartholomew, D. R.; Stauffer, D. B.; McLafferty, F. W. Weighted File Ordering for Fast Matching of Mass Spectra against a Comprehensive Data Base. *Anal. Chem.* **1981**, *53* (12), 1938–1939. https://doi.org/10.1021/ac00235a051.
(21) 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* (4), 899–903. https://doi.org/10.1021/ac00281a028.
(22) Agilent Technologies. GC/MSD Libraries for the MSD ChemStation. Technical Note. **2000**, 0–2.
(23) Pesyna, G. M.; Venkataraghavan, R.; Dayringer, H. E.; McLafferty, F. W. Probability Based Matching System Using a Large Collection of Reference Mass Spectra. *Anal. Chem.* **1976**, *48* (9), 1362–1368. https://doi.org/10.1021/ac50003a026.
(24) Mclafferty, F. W.; Stauffer, B.; Zhang, M.; Loh, S. Y. Comparison of Algorithms and Databases for Matching Unknown Mass Spectra. *J. Am. Soc. Mass Spectrom.* **1998**, *9*, 92–95.
(25) Koo, I.; Zhang, X.; Kim, S. Wavelet- and Fourier-Transform-Based Spectrum Similarity Approaches to Compound Identification in Gas Chromatography/Mass Spectrometry. *Anal. Chem.* **2011**, *83* (14), 5631–5638. https://doi.org/10.1021/ac200740w.
(26) 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* (8), 1158–1163. https://doi.org/10.1093/bioinformatics/bts083.
(27) NIST. NIST Mass Spectral Database for NIST/EPA/NIH and Mass Spectral Search Program (Version 2.3). *Natl. Inst. Stand. Technol. NIST* **2017**, No. Nist 17, 1–73.
(28) Bonetti, J. Mass Spectral Differentiation of Positional Isomers Using Multivariate Statistics. *Forensic Chem.* **2018**, *9*, 50–61. https://doi.org/10.1016/j.forc.2018.06.001.
(29) 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* (July), 77–86. https://doi.org/10.1016/j.forc.2018.10.002.
(30) Bodnar Willard, M. A.; Smith, R. W.; McGuffin, V. L. Statistical Approach to Establish Equivalence of Unabbreviated Mass Spectra. *Rapid Commun. Mass Spectrom.* **2014**, *28* (1), 83–95. https://doi.org/10.1002/rcm.6759.
(31) 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. https://doi.org/10.1016/j.forc.2018.07.001.
(32) Jackson, G. P.; Mehnert, S. A.; Davidson, J. T.; Lowe, B. D.; Ruiz, E. A.; King, J. R. Expert Algorithm for Substance Identification Using Mass Spectrometry: Statistical Foundations in Unimolecular Reaction Rate Theory. *J. Am. Soc. Mass Spectrom.* **2023**, 34(7), 1248-1262. https://doi.org/10.1021/jasms.3c00089.
(33) Sleno, L.; Volmer, D. A. Ion Activation Methods for Tandem Mass Spectrometry. *J. Mass*

Spectrom. **2004**, *39* (10), 1091–1112. https://doi.org/10.1002/jms.703.
(34) Rosenstock, H. M.; Wallenstein, M. B.; Wahrhaftig, A. L.; Eyring, H. Absolute Rate Theory for Isolated Systems and the Mass Spectra of Polyatomic Molecules. **1952**, *38*, 667–678.
(35) Nishimura, T. Fundamentals of Mass Spectrometry. *Fundam. Mass Spectrom.* **2013**, *9781461472*, 1–239. https://doi.org/10.1007/978-1-4614-7233-9.
(36) Samokhin, A. Spectral Skewing in Gas Chromatography–Mass Spectrometry: Misconceptions and Realities. *J. Chromatogr. A* **2018**, *1576*, 113–119. https://doi.org/10.1016/j.chroma.2018.09.033.
(37) Rice, O. K.; Ramsperger, H. C. Theories of Unimolecular Gas Reactions at Low Pressures. *J. Am. Chem. Soc.* **1928**, *50* (3), 617–620. https://doi.org/10.1021/ja01390a002.
(38) Rice, O. K.; Ramsperger, H. C. Theories of Unimolecular Gas Reactions at Low Pressures II. **1928**, *1617* (1927).
(39) Bauer, C. A.; Grimme, S. How to Compute Electron Ionization Mass Spectra from First Principles. *J. Phys. Chem. A* **2016**, *120* (21), 3755–3766. https://doi.org/10.1021/acs.jpca.6b02907.
(40) Marcus, R. A. Unimolecular Dissociations and Free Radical Recombination Reactions. *J. Chem. Phys.* **1952**, *20* (3), 359–364. https://doi.org/10.1063/1.1700424.
(41) Marcus, R. A.; Rice, O. K. Session on Free Radicals the Kinetics of the Recombination of Methyl Radicals and Iodine Atoms. *J. Phys. Colloid Chem.* **1951**, *55* (6), 894–908. https://doi.org/10.1021/j150489a013.
(42) Vékey, K. Internal Energy Effects in Mass Spectrometry. *J. Mass Spectrom.* **1996**, *31* (5), 445–463. https://doi.org/10.1002/(SICI)1096-9888(199605)31:5<445::AID-JMS354>3.0.CO;2-G.
(43) Mehnert, S. A.; Davidson, J. T.; Adeoye, A.; Lowe, B. D.; Ruiz, E. A.; King, J. R.; Jackson, G. P. Expert Algorithm for Substance Identification Using Mass Spectrometry: Application to the Identification of Cocaine on Different Instruments Using Binary Classification Models. *J. Am. Soc. Mass Spectrom.* **2023**, 34(7), 1235-1247. https://doi.org/10.1021/jasms.3c00090.
(44) Bodnar Willard, M. A.; McGuffin, V. L.; Smith, R. W. Statistical Comparison of Mass Spectra for Identification of Amphetamine-Type Stimulants. *Forensic Sci. Int.* **2017**, *270*, 111–120. https://doi.org/10.1016/j.forsciint.2016.11.013.
(45) Bayat, P.; Lesage, D.; Cole, R. B. Tutorial: Ion Activation in Tandem Mass Spectrometry Using Ultra-High Resolution Instrumentation. *Mass Spectrom. Rev.* **2020**, *39* (5–6), 680–702. https://doi.org/10.1002/mas.21623.
(46) 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* (24), 3731–3736. https://doi.org/10.1002/rcm.2247.
(47) Samokhin, A.; Sotnezova, K.; Lashin, V.; Revelsky, I. Evaluation of Mass Spectral Library Search Algorithms Implemented in Commercial Software. *J. Mass Spectrom.* **2015**, *50* (6), 820–825. https://doi.org/10.1002/jms.3591.
(48) Carby-Robinson, D.; Dalsgaard, P. W.; Mollerup, C. B.; Linnet, K.; Rasmussen, B. S. Cocaine Profiling Method Retrospectively Developed with Nontargeted Discovery of Markers Using Liquid Chromatography with Time-of-Flight Mass Spectrometry Data. *Drug Test. Anal.* **2021**, No. March, 1–12. https://doi.org/10.1002/dta.3130.
(49) Risum, A. B.; Bro, R. Using Deep Learning to Evaluate Peaks in Chromatographic Data.

Talanta **2019**, *204* (May), 255–260. https://doi.org/10.1016/j.talanta.2019.05.053.
(50) Ruiz, E. A.; Davidson, T. J.; Jackson, G. P. *Identifying the Sources of Variance of Ion Abundances of GC-EI-MS Measurements*; 2018.
(51) Casale, J. F. *The Mass Spectrum of Cocaine: Deuterium Labeling and MS/MS*; 2010.
(52) Allen, A. C.; Cooper, D. A.; Kiser, W. O.; Cottrell, R. C. The Cocaine Diastereoisomers. *J. Forensic Sci.* **1981**, 12–26.

7. Appendix

Table 29. Descriptive statistics for all 27 possible combinations of factors for *m/z* 42.

m/z	Repeller voltage	Ion focus voltage	EI energy	Mean	Std. Deviation	N
42	20	70	65	27.577	1.9244	12
			70	22.747	1.7045	14
			80	19.135	1.7177	18
			Total	22.586	3.8647	44
		90	65	34.189	2.2169	12
			70	30.231	2.3211	15
			80	26.278	1.7084	17
			Total	29.783	3.8003	44
		110	65	34.495	1.5744	11
			70	32.322	1.5283	12
			80	29.777	1.4356	16
			Total	31.891	2.4585	39
	30	70	65	18.927	1.3839	23
			70	20.957	1.3669	22
			80	21.175	1.4871	22
			Total	20.332	1.7297	67
		90	65	22.236	1.3861	26
			70	22.790	1.2411	25
			80	21.321	1.1836	30
			Total	22.068	1.3970	81
		110	65	27.118	1.0262	23
			70	27.434	1.5563	23
			80	26.297	1.1719	25
			Total	26.931	1.3421	71
	40	70	65	19.007	1.6114	22
			70	20.313	1.1476	20
			80	21.736	1.4009	19
			Total	20.285	1.7833	61
		90	65	19.247	0.9546	29
			70	19.901	1.1329	31
			80	20.355	1.2712	29
			Total	19.836	1.2025	89
		110	65	23.407	0.9945	28
			70	23.480	1.2045	29
			80	22.683	1.1121	30

			Total	23.182	1.1546	87

Table 30. Descriptive statistics for all 27 possible combinations of factors for *m/z* 303.

m/z	**Repeller voltage**	**Ion focus voltage**	**EI energy**	**Mean**	**Std. Deviation**	**N**
303	20	70	65	3.086	0.5379	12
			70	3.570	0.6704	14
			80	3.930	0.9112	18
			Total	3.585	0.8108	44
		90	65	3.720	0.4509	12
			70	3.632	0.6380	15
			80	4.612	0.8313	17
			Total	4.034	0.8110	44
		110	65	3.540	0.5375	11
			70	3.984	0.4301	12
			80	4.663	0.7210	16
			Total	4.137	0.7496	39
	30	70	65	7.257	2.3307	23
			70	6.874	2.2110	22
			80	7.513	2.5104	22
			Total	7.215	2.3324	67
		90	65	8.069	2.0944	26
			70	8.349	2.4591	25
			80	9.402	2.9500	30
			Total	8.649	2.5864	81
		110	65	7.761	1.6088	23
			70	7.752	1.7745	23
			80	8.972	2.6182	25
			Total	8.185	2.1203	71
	40	70	65	10.305	3.7389	22
			70	10.533	3.5857	20
			80	10.000	3.5333	19
			Total	10.285	3.5715	61
		90	65	15.198	5.3935	29
			70	15.130	5.5943	31
			80	15.922	6.3536	29
			Total	15.410	5.7357	89
		110	65	13.455	4.6028	28
			70	14.020	4.8801	29
			80	14.678	5.5878	30

		Total	14.065	5.0187	87	

Table 31. Eta squared values for all *m/z* values including total averages and averages by *m/z* values.

m/z	Intercept	Repeller	Ion Focus	EI Energy	Repeller * Ion Focus	Repeller * EI Energy	Ion Focus * EI Energy	Repeller * Ion Focus * EI Energy	Error
41	96.16	0.471	1.066	0.004	0.198	0.230	0.011	0.048	1.810
42	96.52	1.381	1.001	0.109	0.296	0.302	0.011	0.046	0.335
51	97.30	0.565	0.879	0.019	0.354	0.268	0.003	0.025	0.592
55	98.37	0.550	0.246	0.059	0.056	0.143	0.003	0.032	0.545
67	97.07	0.048	0.144	0.015	0.010	0.075	0.006	0.010	2.625
68	98.58	0.314	0.200	0.095	0.068	0.119	0.004	0.006	0.613
77	98.32	0.583	0.458	0.108	0.153	0.141	0.009	0.004	0.220
81	99.32	0.032	0.004	0.005	0.004	0.005	0.001	0.001	0.624
82	100	0	0	0	0	0	0	0	0
83	99.08	0.046	0.008	0.000	0.001	0.005	0.002	0.004	0.857
94	99.02	0.096	0.286	0.078	0.105	0.022	0.007	0.004	0.383
96	99.12	0.031	0.047	0.002	0.007	0.002	0.001	0.004	0.791
97	98.60	0.082	0.019	0.010	0.006	0.022	0.004	0.004	1.257
105	99.09	0.046	0.412	0.071	0.065	0.026	0.003	0.009	0.275
122	93.61	2.361	0.141	0.045	0.035	0.032	0.006	0.006	3.764
152	84.83	5.113	0.306	0.072	0.073	0.069	0.013	0.016	9.508
182	91.70	5.310	0.315	0.067	0.020	0.056	0.004	0.008	2.514
183	91.75	5.371	0.386	0.073	0.011	0.086	0.004	0.006	2.312
198	89.29	6.722	0.254	0.062	0.036	0.064	0.007	0.011	3.549
272	76.59	12.26	0.945	0.125	0.373	0.029	0.009	0.013	9.660
303	70.29	14.47	0.969	0.128	0.757	0.017	0.033	0.017	13.32
Average	93.73	2.792	0.404	0.057	0.131	0.086	0.007	0.014	2.778
Average high *m/z*	85.44	7.372	0.474	0.082	0.186	0.051	0.011	0.011	6.375
Average low *m/z*	97.33	0.555	0.589	0.050	0.164	0.189	0.006	0.028	1.087
Average middle *m/z*	98.87	0.147	0.205	0.045	0.056	0.036	0.004	0.005	0.630