

# Cocaine profiling method retrospectively developed with nontargeted discovery of markers using liquid chromatography with time-of-flight mass spectrometry data

Daniel Carby-Robinson  | Petur Weihe Dalsgaard  | Christian Brinch Mollerup  | Kristian Linnet | Brian Schou Rasmussen

Section of Forensic Chemistry, Department of Forensic Medicine, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark

## Correspondence

Daniel Carby-Robinson, Section of Forensic Chemistry, Department of Forensic Medicine, Faculty of Health and Medical Sciences, University of Copenhagen, Frederik V's Vej 11, Copenhagen 2100, Denmark.  
Email: daniel.carby-robinson@sund.ku.dk

## Abstract

Illicit drug profiling performed by forensic laboratories assists law enforcement agencies through providing information about chemical and/or physical characteristics of seized specimens. In this article, a model was developed for the comparison of seized cocaine based on retrospective analysis of data generated from ultrahigh performance liquid chromatography with time-of-flight mass spectrometry (UHPLC-TOF-MS) comprehensive drug screening. A nontargeted approach to discover target compounds was employed, which generated 53 potential markers using data from cocaine positive samples. Twelve marker compounds were selected for the development of the final profiling model. The selection included a mixture of commonly used cocaine profiling targets and other cocaine-related compounds. Combinations of pretreatments and comparison metrics were assessed using receiver operating characteristic curves to determine the combination with the best discrimination between linked and unlinked populations. Using data from 382 linked and 34,519 unlinked distances, a classification model was developed using a combination of the standardization and normalization transformations with Canberra distance, resulting in a linked cut-off with a 0.5% false positive rate. The present study demonstrates the applicability of retrospectively developing a cocaine profiling model using data generated from UHPLC-TOF-MS nontargeted drug screening without pre-existing information about cocaine impurities. The developed workflow was not specific to cocaine and thus could potentially be applied to any seized drug in which there are both sufficient data and impurities present.

## KEYWORDS

chemometrics, cocaine profiling, high-resolution mass spectrometry, retrospective analysis

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

© 2021 The Authors. *Drug Testing and Analysis* published by John Wiley & Sons Ltd.

## 1 | INTRODUCTION

Illicit drug profiling performed by forensic laboratories assists law enforcement agencies through providing information about chemical and/or physical characteristics of seized specimens.<sup>1</sup> This drug profiling is performed using statistical comparison of the chemical profiles of different seized samples of an illicit substances, allowing each pair of samples to be classified as related or different.<sup>2-4</sup> Confirming a suspected relationship, or identifying previously unknown relations between drug seizures, can provide law enforcement agencies with strategic and operational knowledge, allowing them to better combat illegal drug manufacture and trafficking.<sup>1,3,5-7</sup> In 2017 alone, 4872 cocaine seizures were made in Denmark (5.6 million inhabitants), and a total of 150.7-kg material was seized by law enforcement.<sup>8</sup> These values suggest that illegal drug trafficking is a major problem, thereby requiring advancements in tools and methods for providing scientific support to aid in combatting the issue.

Profiling of cocaine samples most often focuses on the chemical profiles resulting from occluded residual solvents, coca leaf derived alkaloids, trace elements or stable carbon and nitrogen isotopes.<sup>1,2,9</sup> The types of chemical profile originate from different sources and will therefore provide different definitions of what can be considered related or linked. Thus, it is important to establish what type of question needs answering to maximize the scientific value of the developed profiling method. In this project, a method for impurity profiling for the batch comparison of cocaine is developed based on alkaloid compounds. During cocaine manufacture, the leaves of the *Erythroxylum coca* plant are processed in clandestine laboratories with crude and unsophisticated production conditions to extract the cocaine, and consequently, the purification of the cocaine is inadequate resulting in many major and minor plant alkaloids remaining present in the final refined drug product.<sup>9-11</sup> Fortunately for the forensic chemist, relative measurements of these alkaloids vary between batches of cocaine, allowing impurity profiling to determine whether two seized cocaine samples originate from the same batch.<sup>1,6,7,12</sup> The origination from the same production batch is considered a 'link' between seizures, whereas two unrelated seizures are considered 'unlinked.' This concept of linked and unlinked is dependent on how the analyte in question is produced and the purpose of the investigation. For the purposes of this study, this definition is considered sufficient to allow us to develop an adequate method for impurity profiling of cocaine seizures.

The impurity profiling of cocaine is traditionally performed using score-based classification models based on data collected from various common analytical techniques.<sup>1</sup> Firstly, peak areas for each marker of the chemical profile are statistically pretreated to account for analytical run variation. Next, the degree of similarity between two or more samples is evaluated using a suitable distance/similarity metric, and finally, the compared samples can be classified as either similar or dissimilar using appropriately similarity/dissimilarity thresholds. For profiling based on alkaloids, the most common analytical technique is gas chromatography with mass spectrometry or nitrogen phosphorous detection (GC-MS or NPD).<sup>1,2,7,12</sup> High-resolution MS

(HRMS) has become more widely employed in forensic chemistry for comprehensive nontargeted screening of seizures and biological matrices for a variety of purposes.<sup>13-17</sup> Full spectrum acquisition allows for sensitive detection of thousands of compounds with a high mass accuracy and resolution. This allows for screening of compounds in suspected cocaine seizures without prior need to preselect target compounds, allowing for a wealth of information to be generated regarding the profile of the impurities.

In this paper, a profiling method was developed for the batch comparison of seized cocaine using retrospective data from comprehensive drug screening using ultrahigh performance liquid chromatography with time-of-flight MS (UHPLC-TOF-MS). To maximize the profiling potential of the method, a novel approach to marker discovery was used. Inspired by workflows used for nontargeted analysis in the field of metabolomics, this 'omics style' approach generated a set of cocaine marker compounds, including both commonly used profiling alkaloids and unidentified compounds.<sup>13,18,19</sup> Using these markers as a foundation, the main objective of the paper was to develop a profiling model for the batch comparison of cocaine samples, which could determine all the linked cocaine samples to be linked (true positive, TP), whilst linking no unlinked samples (true negative, TN).

## 2 | MATERIALS AND METHODS

### 2.1 | Samples

The data for the nontargeted discovery of marker compounds consisted of UHPLC-TOF-MS screening data from analysis of 1962 seized narcotic samples submitted to the authors' laboratory between September 2017 and January 2020. After October 2019, each analysis run included reanalysis of two seized cocaine samples used as cocaine quality control (QC) samples. In total, 1962 single injections in 156 runs had been performed, each containing system samples, blanks and QCs. The data were split based on the determined cocaine concentration into a positive group when the determined cocaine concentration exceeded 5% (w/w) and a negative group if cocaine was not present. If cocaine concentration was below 5%, the seized sample was excluded from further analysis. The marker discovery was performed using retrospective analysis of data from single injections of the positive group samples ( $n = 487$ ) and of the negative group samples ( $n = 1462$ ). Evaluation of marker compounds was performed through retrospective analysis of data from single injections of the positive group samples ( $n = 487$ ) and repeat injections of one cocaine QC sample ( $n = 17$ ).

Profiling method development was performed using data from injections of the positive group samples, both from original screening injections ( $n = 487$ ) and from reanalysis injections ( $n = 20$ ) of five positive samples, in addition to repeat injections ( $n = 43$ ) of cocaine QC samples. For profiling development and evaluation, the samples received for different cases were considered unlinked, whereas all repeated injections of a sample were considered linked.

## 2.2 | Chemicals and standards

LC-MS grade methanol, water, formic acid and acetonitrile were obtained from Fisher Scientific (Loughborough, UK). Leucine enkephalin and carbamazepine were obtained from Sigma-Aldrich (Merck, Darmstadt, Germany).

## 2.3 | Sample preparation

Approximately 50 mg of cocaine sample was weighed into 20 mL plastic centrifuge tube and dissolved in 5 mL of methanol adjusted accordingly, in order to reach a final dilution of 10 mg/mL. Two glass beads were added to each sample, which were placed in a rotation mixer for 10 min. Following this, tubes were centrifuged for 5 min. Samples were then filtered through 0.5- $\mu$ m PTFE filters and diluted 500 times with internal standard solution containing carbamazepine (0.5 mg/L) and 25% methanol in water with 1% formic acid and then transferred to appropriate glass vials ready for injection. The two cocaine QCs were also prepared following the same procedure as the cocaine samples.

## 2.4 | Instrumentation

Nontargeted data acquisition was performed using an UHPLC-TOF-MS consisting of an ACQUITY I-Class UPLC System (Waters Corporation, Milford, MA, USA) coupled to a Xevo G2-S QTOF (Waters). All instrument control, data acquisition and initial data processing were performed with UNIFI Scientific Information Systems (Waters). Chromatographic separation was performed using a C18 column (ACQUITY UPLC HSS C18 Column, 100 Å, 1.8  $\mu$ m, 2.1  $\times$  150 mm; PN: 186003534; Waters) maintained at a column temperature of 50°C with a flow rate of 0.4 mL/min, using 1- $\mu$ L injection volume. Mobile phase A consisted of 5-mM aqueous ammonium formate buffer adjusted to a pH 3 with formic acid. Mobile phase B consisted of acetonitrile with 0.1% v/v formic acid. The gradient programme started with a brief hold at 13% (B) from 0 to 0.5 min, followed by a linear ramp of 13–50% from 0.5 to 10 min. Next, mobile phase B was further increased from 50% to 95% from 10 to 10.75 min, followed by a final hold at 95% from 10.75 to 12.25 min. The chromatographic programme finished with equilibration from 12.25 min to a final runtime of 15 min.

The TOF-MS was operated in positive electrospray ionization mode (Z-Spray; Waters) with the following settings: 1000 L/h of nitrogen as the nebulization gas at 400°C, 20 L/h cone gas flow with at 150°C, a capillary voltage of 800 V, a cone voltage of 25 V and argon as the collision gas. Data were recorded in profile mode using the data independent acquisition mode: MS<sup>E</sup>. The low collision energy (CE) was set at 4 eV, and high CE ramped from 10 to 40 eV. The acquisition time spanned the whole run, with a scan time of 0.200 s and a mass range from  $m/z$  50 to 950. Weekly mass calibration was

performed with 5-mM sodium formate in propanol and water (90:10, v:v). Leucine enkephalin ( $m/z$  556.2766) was acquired every 30 s from a reference spray, for use as lock mass correction using three consecutive scans.

## 2.5 | Data processing

Feature detection was performed using the UNIFI 3D-peak detection with intensity thresholds of five counts in the low energy channel and the high energy channel, with noise background filtering set to high, and a total number of peaks to keep per channel set to 10,000,000. No target components were included in the method. The analysis data were exported as UNIFI export package (uep) files. Feature lists from the uep files were read and imported into a local SQL database (SQL Server 2019, Microsoft, Redmond, WA, USA) using scripts developed in Python 3. All further data treatment and statistical work were performed using open-source python packages with Python 3.7.4. The main python packages used for data treatment and statistical work were pandas, NumPy, SciPy and Scikit-learn.<sup>20–23</sup> Figures were made using seaborn and matplotlib.<sup>24,25</sup> Code development and further data analysis were performed using Jupyter notebooks with JupyterLab.

## 2.6 | Nontargeted workflow for marker compound discovery

A nontargeted marker discovery workflow, modified from a previously published 'omics-based' workflow, was employed to enable discovery of potentially novel markers for cocaine profiling. The workflow was completed in four steps: (1) extraction of peaks, (2) grouping of peaks into potential markers, (3) scoring of peak groups based on presence in cocaine seizure samples and, finally, (4) targeted extraction and evaluation of markers for profiling suitability.

### 2.6.1 | Extraction of peaks and grouping of peaks into potential markers (steps 1 and 2)

Peaks, that is, the UNIFI 3D-peak detected features, eluting before 11 min, were extracted from all injections of the positive and negative sample groups using a lower intensity threshold of 200 counts. The extracted peaks were clustered into groups using the mean-shift clustering algorithm available from the Scikit-learn python package.<sup>23</sup> First, all peaks were scaled using retention time and  $m/z$  tolerances: 0.2 min and  $m/z$  0.003. The mean-shift clustering algorithm was then applied using a bandwidth of 1. The values of the newly identified peak groups were then rescaled using the selected tolerances, giving the corresponding peak group for each peak, with the peak group retention time and  $m/z$  values being the average of the peaks contained within the group.

### 2.6.2 | Scoring of peak groups based on presence in cocaine seizure samples (step 3)

The list of peak groups was reduced to those relevant for cocaine profiling using the Matthews correlation coefficient (MCC).<sup>26</sup> As with most evaluation scores, the MCC was calculated based on the values in the confusion matrix, that is, the number of TP, false positive (FP), TN and false negative (FN).<sup>26</sup>

The presence and absence of a peak in a sample were tallied per peak group per sample group, with TP and TN counts representing the number of times the presence or absence of a peak was observed in the positive sample group, respectively. Conversely, FP and FN counts represented the number of times the presence or absence of a peak was observed in the negative sample group respectively.

Based on these values, the MCC was calculated for each peak group using the following equation:

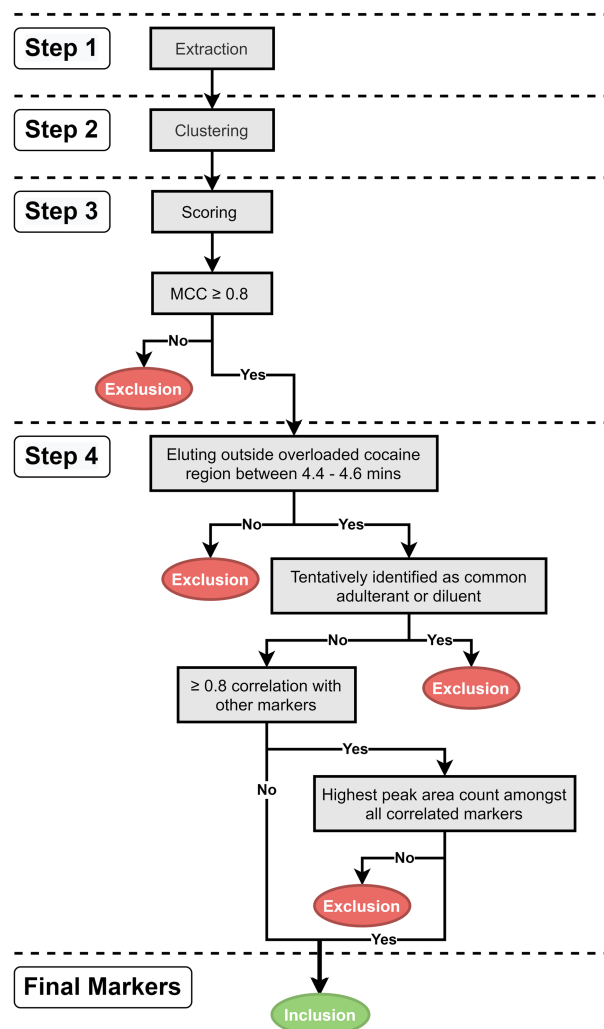
$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FN)(TP + FP)(TN + FP)(TN + FN)}}$$

The MCC is a number between  $-1$  and  $1$ , with  $1$ ,  $0$  and  $-1$  indicating fully accurate, random and fully inaccurate predictions, respectively. Using an MCC cut-off of  $0.8$ , the peak groups were reduced to a list of potential markers present in the majority of the positive group samples, whilst being absent from the majority of the negative group samples.

### 2.6.3 | Targeted extraction and evaluation of markers for profiling suitability (step 4)

The potential markers were used for targeted extraction of the 500 cocaine injections to be used for subsequent evaluation of markers. Evaluation was performed according to the workflow in Figure 1, in order to exclude unsuitable marker compounds from the final marker list. Firstly, markers co-eluting with cocaine in the retention time region of  $4.4$ – $4.7$  min were excluded. The remaining markers were identified using the routine screening library.

The list of potential marker compounds was further reduced by exclusion of highly correlated markers. To achieve this, agglomerative hierarchical clustering with complete-linkage clustering and Pearson's correlation coefficient (PCC) was used with data of the peak area counts for the potential markers across the 500 cocaine injections. Agglomerative hierarchical clustering is an unsupervised machine learning technique for building a hierarchy of clusters.<sup>23</sup> The PCC is calculated between all pairs of potential markers, forming a distance matrix. The two markers with the shortest distance, that is, the most correlated, were then formed into a single cluster. The distance matrix was then recalculated, using the largest of the distances in the new cluster. This process was repeated until the shortest distance between markers was below  $0.8$  PCC. A PCC threshold of  $0.8$  was selected to determine the number of clusters, as a PCC between  $0.8$  and  $1.0$  indicates a strong positive correlation between pairs of markers. In the



**FIGURE 1** Decision tree assessing the suitability of target compounds for inclusion during steps 3 and 4 in the development of the final profiling model

resulting clusters, the mean peak area across the 483 cocaine positive samples was calculated for each potential marker. The marker that presented the largest mean peak area within their respective clusters was included in the final marker list, whereas all other markers were excluded.

Variability was assessed across repeat injections ( $n = 17$ ) of a single cocaine QC, and single injections of all cocaine positive samples ( $n = 483$ ) were determined for each potential marker. The mean relative standard deviation (RSD) was calculated from the peak area of each potential marker across the repeat injections of the cocaine QC and all cocaine samples.

## 2.7 | Profiling method development and evaluation

The determination of whether a sample was linked with another sample and was performed using a comparison metric with decision

thresholds for linked, uncertain and not linked. Prior to use of a comparison metric, the data underwent scaling and transformation, that is, pretreatment, to reduce the instrumental influences on the data sets, that is, the uncontrolled variation not pertaining to the analytical information.<sup>1</sup> The development and evaluation of the profiling model were completed in five steps: (1) selection of the linked and unlinked pairwise distances, (2) calculations of the combinations of pretreatments and comparison metrics, (3) calculations of the area under the curve (AUC) of the receiver operating characteristic (ROC) curves, (4) determination of the decision thresholds and, finally, (5) an evaluation of the profiling model.

### 2.7.1 | Selection of linked and unlinked populations (step 1)

Two populations of pairwise distances were created using the distances between the linked and unlinked cocaine samples of the positive group samples. The linked pairs composed of 382 pairwise distances calculated from all possible pairs of repeat injections for each of the cocaine QCs ( $n = 356$ ), and all pairs of four repeat injections each of the same five cocaine samples ( $n = 26$ ). Unlinked pairs comprised 34,159 pairwise distances calculated from all possible pairs of 483 cocaine positive samples pertaining to different police cases. When multiple samples were received pertaining to a single case, these comparisons were ignored.

### 2.7.2 | Combinations of pretreatments and comparison metrics (step 2)

A combinatory chemometric approach was employed to determine the most suitable combinations of pretreatments and comparison metrics. Four pretreatments and five comparison metrics had been shown by others to be the most suitable when developing a profiling method.<sup>1,7,27-29</sup> Fourteen total variations of these pretreatments (Table S1) with the five comparison metrics (Table S2) were tested for a total of 70 combinations.

### 2.7.3 | Selection of the model giving highest separability (step 3)

The separability was determined using the AUC of the ROC curves. The ROC curves, calculated using the Scikit-learn python package, depicted the TP rate, that is, the proportion of correctly linked samples, as a function of the FP rate, that is, the proportion of falsely linked samples, for the classification models, that is, the 70 profiling methods, with two distinct classes, that is, linked or unlinked samples<sup>30,31</sup>. The AUC of each ROC curve then represented the degree of separability between the linked and unlinked populations and was therefore used as a metric for evaluating the performance of the different combinations.<sup>6,30</sup> To have an AUC value of 1, a model would

have no falsely linked nor falsely unlinked populations and thereby have successfully determined all linked samples as linked and vice versa. The combination that gave the highest AUC was chosen for the final method.

### 2.7.4 | Determination of decision thresholds (step 4)

The decision thresholds were calculated using the information generated from the ROC curves and represented the distance boundaries for samples to be considered linked, inconclusive or unlinked. The decision thresholds were chosen using a continuous probability distribution histogram plot, created using the matplotlib python package.<sup>25</sup> Decision thresholds were selected at two distances in the probability distribution in which the FP and FN rates were 0.5%.

### 2.7.5 | Evaluation of the profiling model (step 5)

Finally, the model was employed through creation of a hierarchical clustered heatmap in order to visualize the hierarchical clustering of the cocaine samples and observe undiscovered links across samples thought to be unlinked. Agglomerative hierarchical clustering was performed using complete linkage and transformed peak area counts for the final marker list on the 483 single injections of the cocaine positive group. Clusters were determined at the previously calculated decision threshold linked samples.

## 3 | RESULTS AND DISCUSSION

### 3.1 | Nontargeted workflow for marker compound discovery

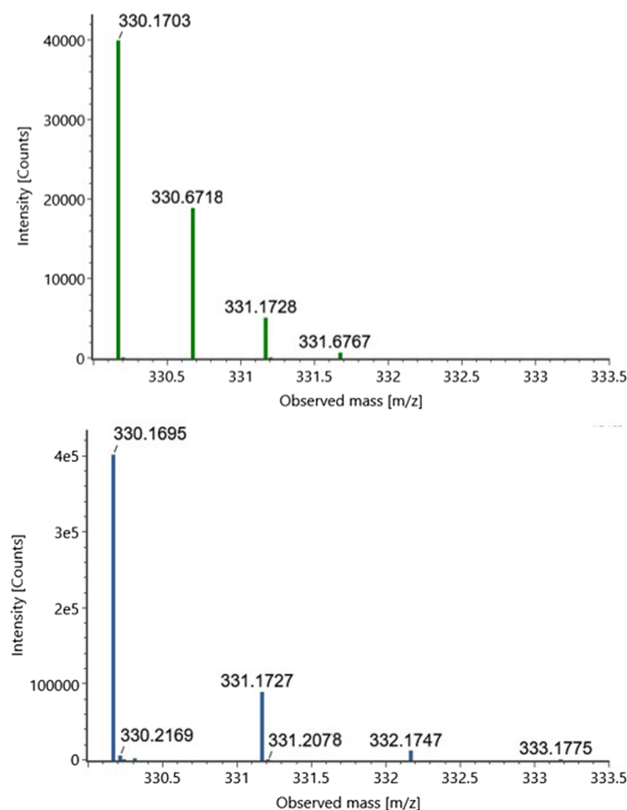
The ideal markers to be used for the development of the cocaine profiling method should be present due to the cocaine content and preferably robust, that is, not be influenced by any treatment of the cocaine samples, such as dilution by addition of adulterants. In this article, an omics-based approach was utilized for nontargeted discovery of all potential targets and filtering of irrelevant variables from the retrospective screening data. The targets initially were discovered using the observed differences between sample groups without pre-existing knowledge, similar to the workflow employed in the metabolomics and other omics fields.

In the first step of the nontargeted marker discovery process, 1,605,653 peaks were extracted from the suitable retention time range (0–11 min) of the nontargeted discovery data set ( $n = 1962$ ) using a peak area count threshold of 200. In Step 2, grouping was performed using tolerances of 0.2 min and  $m/z$  0.003, forming 63,854 peak groups. The mean-shift algorithm was chosen, as this allows for grouping of the peaks across injections, accounting for shifts in retention time and the accurate mass measurements, whilst not requiring

knowledge regarding of the total number of groups. In Step 3, the MCC was calculated, and 78 peak groups were selected using a MCC threshold of 0.8.

The MCC was calculated based on the confusion matrix of observations in the positive and negative sample set. It is used to assess the quality of a binary classifier, and to get a high MCC score, a peak group had to be present in the majority of the positive group samples and absent in the majority of the negative group samples, independent of the positive and negative ratio of the dataset. The inclusion of cocaine negative samples and the use of a high MCC threshold allowed for efficient removal of peaks likely to originating from another source than the cocaine, that is, peaks originating from adulterants or contaminants. Adulterants can be added to a cocaine seizure at any stage of the cocaine production and distribution chain, therefore are not relevant for batch comparison of cocaine.<sup>32–34</sup> To calculate the confusion matrix, the presence or absence of the classifier needs to be determined; that is, the peak is observed or not observed. However, because the peaks were extracted without use of any targets, no inherent grouping was available.

The now 78 peak groups could be considered potential markers, and peak count information was extracted from the 500 injections of cocaine positive samples ( $n = 483$ ) and QCs ( $n = 17$ ) to be used for marker evaluation. The initial abundance of alkaloid impurities sourced from cocaine manufacture is proportionate to the concentration of the cocaine in the seizure; therefore, cocaine samples of 5% concentration or less were excluded from the positive sample group. An evaluation was performed using the workflow in Figure 1, to ensure selection of quality markers, to consequently reduce the number of FP or FN associations in the final profiling model. Firstly, 25 markers co-eluting with the overloaded cocaine were excluded, as they eluted within the retention time range of 4.4–4.7 min. Most of these markers appeared to be isotopologues or fragments directly resulting from the cocaine, therefore provided as little support for distinguishing seizures as the cocaine itself and consequently were removed. The remaining markers were labelled M1–M53 in order of retention time (Table S3). The markers were identified by comparison with our in-house screening library. Four compounds were identified through this comparison as the alkaloids ecgonine methyl ester, benzoylecgonine, tropacocaine and trans-cinnamoylcocaine. Detection of the four compounds was consistent with previous publications concerning impurity profiles of cocaine seizures.<sup>2,6,7,29,35,36</sup> Elemental compositions for the 12 marker compounds were determined using UNIFI software. For the eight unknown marker compounds without reference standards, identification was performed tentatively through inspection of spectra. Both cinnamoylcocaine and the  $[M + 2H]^+$  ion of truxilline share the same exact mass; therefore in order to distinguish the two, extracted ion chromatograms (EICs) were used. As shown in Figure 2, the double-charged truxilline contains the characteristic double-charged isotope pattern with approximately 0.5-mDa spacing between C13 isotopes. By producing an EIC at 330.6717, markers M23 and M49 that shared the same exact mass were able to be tentatively identified as cis-cinnamoylcocaine and  $[M + 2H]^+$  ion of truxilline, respectively, as shown in Table 1. M6 appeared to be the common profiling alkaloid



**FIGURE 2** Mass spectra highlighting the isotope pattern of a double-charged truxilline (top) and a cinnamoylcocaine (bottom)

3,4,5-trimethoxyocaine based on mass and fragmentation pattern in comparison with previous research. M8, M21, M31, M36 and M46 all were tentatively identified as either the  $[M + H]^+$  or  $[M + 2H]^+$  ions of the demethylated truxilline compound. Around 11 different isomers of truxilline have been reported in cocaine samples and are the result of photo-dimerization of cis/trans-cinnamoylcocaines in the coca leaf.<sup>37</sup> Of the 12 final markers, M1, M2, M4, M6, M23 and M38 were all identified as markers, which have been commonly used for impurity profiling of cocaine, as shown in Table 1. This consensus with other literature provides support for the developed profiling model and marker selection process. In addition, none of the potential markers were tentatively identified as common adulterants, diluents or other drugs of abuse that may have been mixed with cocaine samples, implying the successful removal of peaks originating from such sources using the high MCC score.

The list of markers was further reduced based on the observed correlation between markers. To assess correlation, the PCC treatment was applied to all possible pairs of markers from the cocaine positive sample group. Twelve clusters of correlated markers were formed from the 53 potential markers using a 0.8 correlation threshold, as shown in the dendrogram in Figure 3. The marker with highest mean peak area count across the 500 cocaine positive samples was chosen from each cluster of correlated markers, to ensure the selection of the base peak.

**TABLE 1** Overview of the final 12 compounds selected for development of the profiling model

ID	Compound	RT (min)	Mean shifted <i>m/z</i>	Chemical formula	Ion	<i>m/z</i> error (mDa)	References
M1	Ecgonine methyl ester <sup>+a</sup>	0.80	200.1280	C <sub>10</sub> H <sub>17</sub> NO <sub>3</sub>	[M + H] <sup>+</sup>	0.0001	Previous studies <sup>2,6,7,35–39</sup>
M2	Benzoyllecgonine <sup>+a</sup>	2.94	290.1385	C <sub>16</sub> H <sub>19</sub> NO <sub>4</sub>	[M + H] <sup>+</sup>	0.0002	Previous studies <sup>7,10,35–37</sup>
M4	Tropacocaine <sup>+a</sup>	4.01	246.1487	C <sub>15</sub> H <sub>19</sub> NO <sub>4</sub>	[M + H] <sup>+</sup>	0.0002	Previous studies <sup>2,5,6,10,29,35–39</sup>
M6	3,4,5-Trimethoxycocaine <sup>+</sup>	4.71	394.1856	C <sub>20</sub> H <sub>27</sub> NO <sub>7</sub>	[M + H] <sup>+</sup>	0.0004	Previous studies <sup>2,5–7,29,35–38</sup>
M8	Demethylated truxilline + 2H <sup>+</sup>	4.78	323.1622	C <sub>37</sub> H <sub>44</sub> N <sub>2</sub> O <sub>8</sub>	[M + 2H] <sup>+</sup>	0.0000	
M21	Demethylated truxilline + H <sup>+</sup>	5.31	645.3163	C <sub>19</sub> H <sub>23</sub> NO <sub>4</sub>	[M + H] <sup>+</sup>	0.0007	
M23	Cis-cinnamoylcocaine <sup>+</sup>	5.72	330.1701	C <sub>19</sub> H <sub>23</sub> NO <sub>4</sub>	[M + H] <sup>+</sup>	0.0001	Previous studies <sup>2,5–7,29,35–38</sup>
M31	Demethylated truxilline + H <sup>+</sup>	5.82	645.3163	C <sub>37</sub> H <sub>44</sub> N <sub>2</sub> O <sub>8</sub>	[M + H] <sup>+</sup>	0.0007	
M36	Demethylated truxilline + 2H <sup>+</sup>	6.18	323.1622	C <sub>37</sub> H <sub>44</sub> N <sub>2</sub> O <sub>8</sub>	[M + 2H] <sup>+</sup>	0.0000	
M38	Trans-cinnamoylcocaine <sup>+a</sup>	6.28	330.1700	C <sub>19</sub> H <sub>23</sub> NO <sub>4</sub>	[M + H] <sup>+</sup>	0.0000	Previous studies <sup>2,5–7,29,35–38</sup>
M46	Demethylated truxilline + 2H <sup>+</sup>	6.47	323.1622	C <sub>37</sub> H <sub>44</sub> N <sub>2</sub> O <sub>8</sub>	[M + 2H] <sup>+</sup>	0.0000	
M49	Truxilline + 2H <sup>+</sup>	6.61	330.1701	C <sub>38</sub> H <sub>46</sub> N <sub>2</sub> O <sub>8</sub>	[M + 2H] <sup>+</sup>	0.0001	

<sup>a</sup>Identified via retention time and mass spectral comparison to in-house libraries created from reference standards.

The list of markers contained multiple observations of the same underlying compound, that is, isotopologues and adduct peaks of the same compound. Whilst isotopologues are naturally present, occurrence of metal adduct ions, such as [M + Na]<sup>+</sup> and [M + K]<sup>+</sup>, is expected with use of positive electrospray ionization. These adducts can originate from solvent impurities, instrument conditions or mobile phase additives.<sup>40</sup> Isotopologues and adduct peaks can be elucidated using the observed mass differences between co-eluting peaks; however, no such approach was found necessary, as the high correlation between isotopologues and subsequent removal of all but the base peak allowed for similar peak reduction.

Two of the established profiling alkaloids, ecgonine methyl ester (M1) and benzoyllecgonine (M2), did not appear to correlate strongly with any other compounds, forming individual clustering and thus were included in the final set of profiling markers. The other two of the identified alkaloids, trans-cinnamoylcocaine (M35) and tropacocaine (M3), formed cluster pairs with their associated <sup>13</sup>C isotope peaks and as a result presented a higher peak area and were thereby also included in the final set of profiling markers. Markers M23 and M25 also formed a cluster pair containing a base peak and associated <sup>13</sup>C isotope, with the M23 determined as the base peak based on the mean peak count. The remaining 45 markers were clustered into seven correlated peak groups, resulting in a total of 12 cocaine profiling markers, as shown in Table 1.

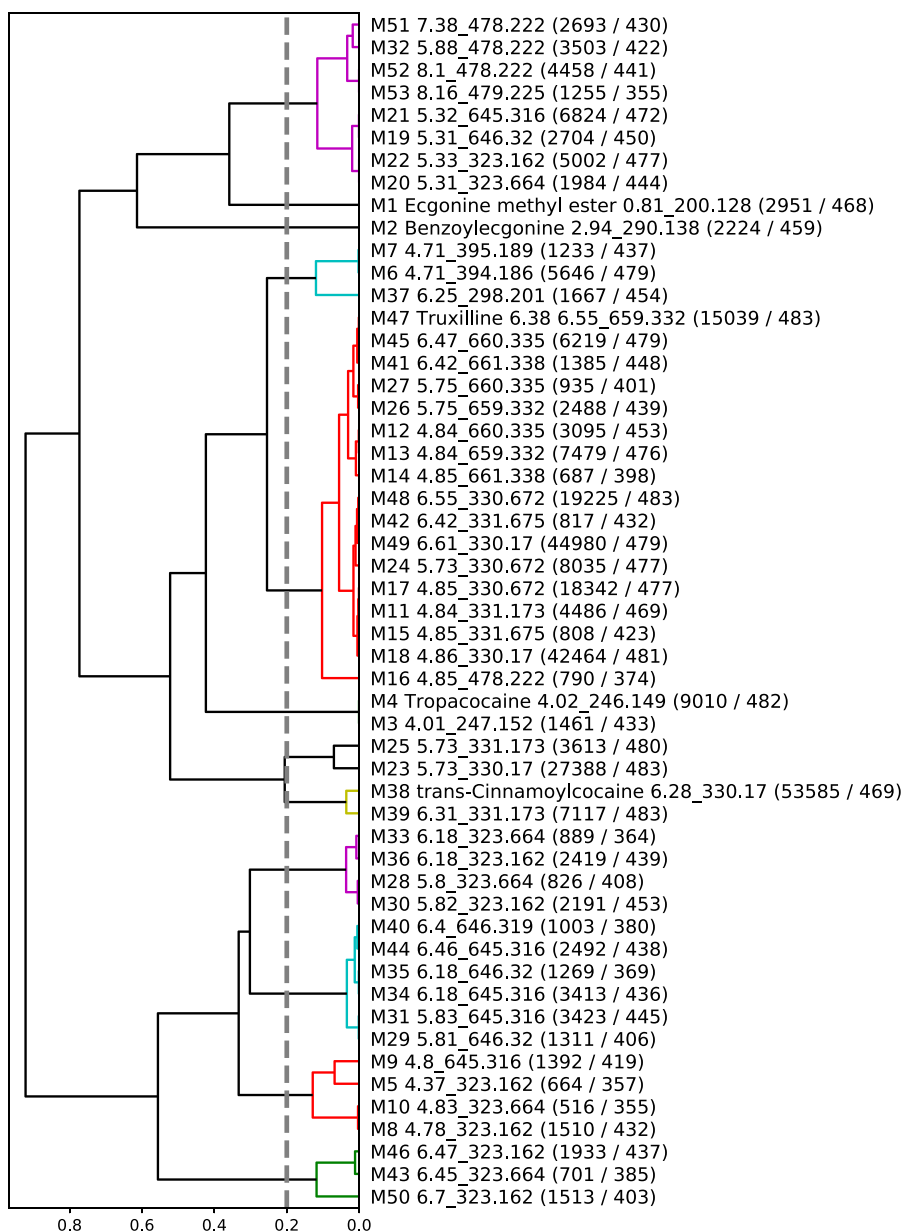
The RSD values for the final 12 compounds (Table 1) ranged from 1.02% to 7.33% (mean: 2.51%) amongst the repeated measures

(*n* = 17) and 15.3% to 51.1% (mean: 24.3%) amongst all single cocaine injections (*n* = 483). This suggested that the final markers appeared relatively stable across repeat measurements, whilst having a high variance in different cocaine samples thereby supporting their suitability as profiling targets. For the purposes of profiling, full identification of all compounds was not necessary; therefore, markers that were previously identified were labelled appropriately with a marker number, retention time and *m/z*. However, many unidentified markers have masses and isotope patterns that fit common alkaloids in cocaine such as truxillines.<sup>41</sup>

### 3.2 | Development of profiling model

The 12 compounds now represent the impurity profile and were used as a foundation to develop a model for the batch comparison of cocaine. In development of profiling models in previous research, linkage amongst cocaine samples is evaluated based on case information.<sup>7,35–38,42</sup> A more objective and reproducible approach was used, with known linked pairs determined from repeat measures of reanalysed cocaine samples and cocaine QCs, requiring no pre-existing knowledge regarding the linkage of seizures. This avoids the tedious and subjective assignment of linkage status based on using information from police case, which can be challenging when developing a method retrospectively. One limitation of this, however, is the lack of inclusion of pairs from authentic 'linked' seizures. The repeat

**FIGURE 3** Dendrogram from hierarchical clustering of the 53 potential markers from the nontargeted marker discovery. Markers are labelled with identification (if known), followed by retention time, mass, the mean peak area across all cocaine positive samples and count in cocaine positive samples (out of 483)



measures may not be representative of a 'link' between authentic seizures, although provides a sufficient alternative when linkage is not possible based on case information. To determine unlinked status, it is more probable that most samples from different cases originate from different production batches. For this reason, all possible pairs of injections of cocaine samples from different police cases were used to constitute the unlinked population.

In Step 1 of the development, the pretreatments and distance measures used in the combinatory approach were chosen based on application and documented usage in profiling models for a wide range of drug substances.<sup>1,28,36,38,43-45</sup> The combinatory approach is also commonly used for developing profiling models, to determine the best separation of the linked and unlinked data. In Step 2 of the development, AUC values ranged from 0.861 to 0.999 for the evaluated combinations, with non-pretreated models giving less discrimination,

which can be seen in Table 2. Many combinations presented an AUC score above 0.98 (Table 2), with the Canberra (CAN) and Manhattan (MAN) comparison metrics appearing to outperform the others based on AUC score alone. The standardization and normalization (S + N) pretreatment with CAN was selected for this final profiling model, with the associated ROC plot shown in Figure 4. To date, combinations involving square cosine function (SCF) and PCC were seemingly favoured for cocaine profiling in similar articles employing a combinatory approach.<sup>7,36,37</sup> However, in this study, the effectiveness of the lower ranked options was not significantly lower, with 55 out of 71 with an AUC greater than 0.95. This demonstrates the robustness of the developed model to changes in comparison metrics and suggests that selection of the optimal combination may not be as important as a parameter for developing the profiling model as the quality of the data or selection of target markers.

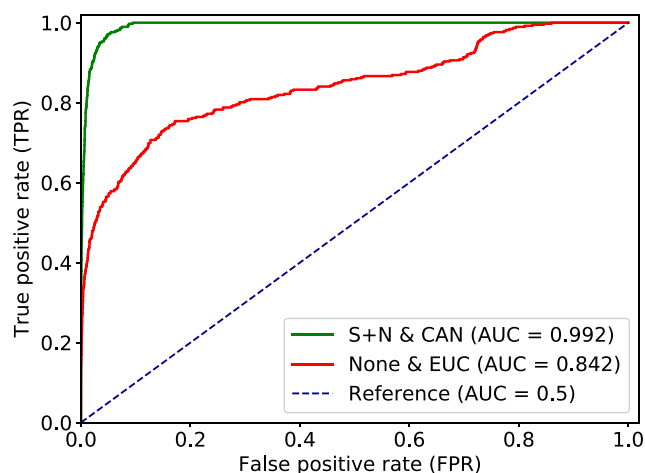


**TABLE 2** AUC values for every comparison metric and pretreatment combination

Pretreatment	Pearson (PCC)	Cosine (SCF)	Euclidean (EUC)	Canberra (CAN)	Manhattan (MAN)
-	0.930	0.932	0.841	0.953	0.856
N	0.930	0.932	0.928	<b>0.988</b>	0.949
S	0.973	0.970	0.910	0.953	0.901
L	0.967	0.970	0.970	0.959	0.960
4R	0.944	0.954	0.957	0.959	0.955
N + S	0.977	0.977	0.969	<b>0.988</b>	0.976
N + L	0.956	0.961	0.969	<b>0.987</b>	0.958
N + 4R	0.944	0.954	0.957	<b>0.987</b>	<b>0.980</b>
N + S + 4R	0.966	0.972	0.976	<b>0.987</b>	<b>0.986</b>
N + S + L	0.951	0.962	0.976	<b>0.981</b>	<b>0.987</b>
S + N	0.973	0.970	0.970	<b>0.992</b>	<b>0.982</b>
S + L	0.965	0.899	0.970	0.835	0.960
S + N + 4R	0.971	0.978	0.978	<b>0.990</b>	<b>0.991</b>
S + N + L	0.973	0.974	0.970	<b>0.990</b>	<b>0.990</b>

Note: Combinations presenting an AUC exceeded 0.98 are highlighted in bold.

Abbreviations: AUC, area under the curve; EUC, Euclidean distance; PCC, Pearson's correlation coefficient; SCF, square cosine function.

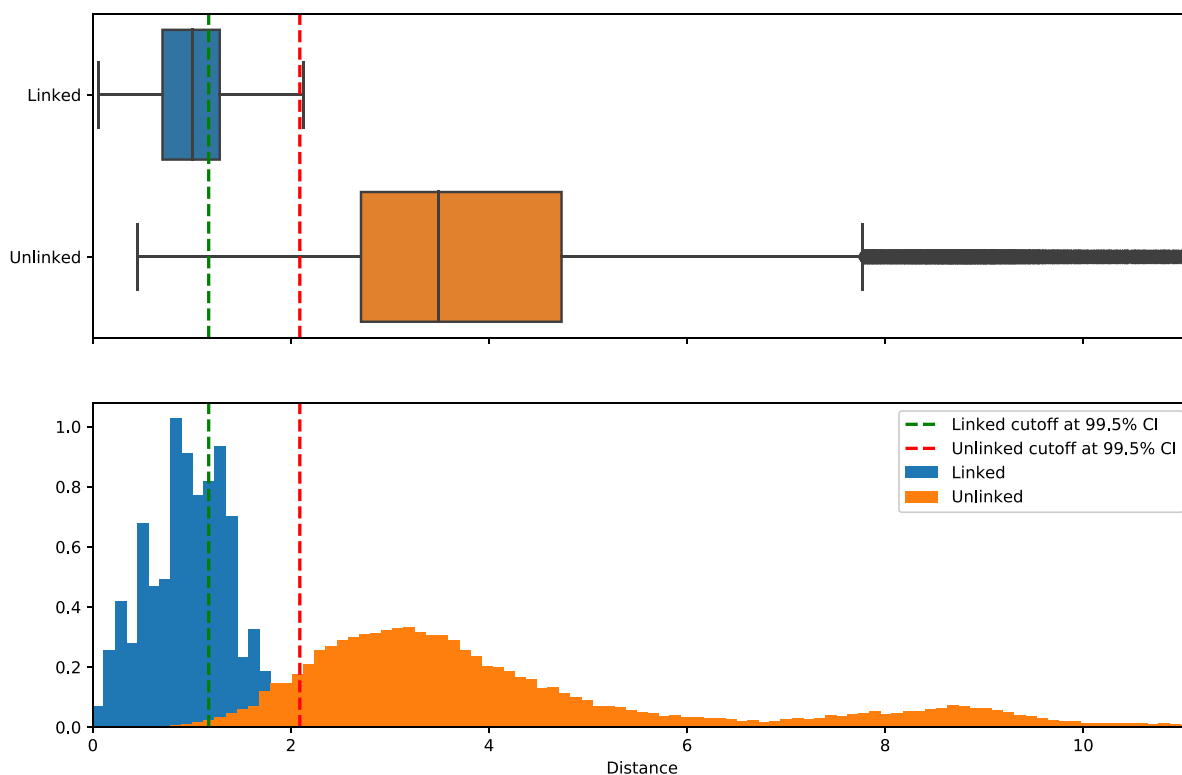


**FIGURE 4** Receiver operating curve (ROC) plot with associated area under the curve (AUC) values, highlighting the difference in performance of two pretreatment and comparison metric combinations. The standardization and normalization treatment with Canberra distance (S + N and CAN) is an example of a better performing comparison metric, whereas the untreated Euclidean distance (EUC) is an example of a worse performing combination

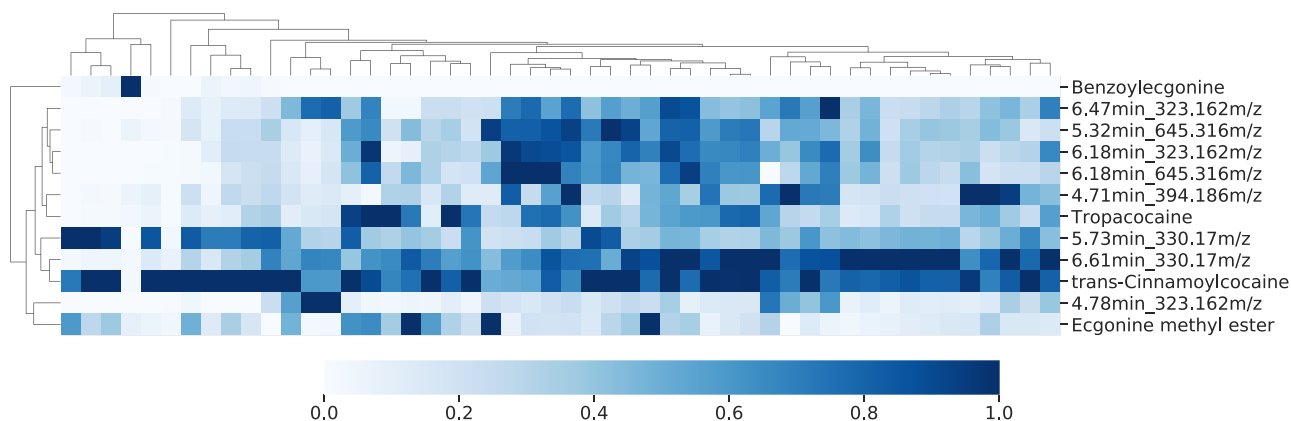
In the next step, suitable thresholds were chosen in order to provide the method with a suitable level of performance for real world application. As shown in the boxplot and histogram in Figure 5, distributions for the unlinked and linked distances overlapped; therefore, the selected decision thresholds were decided based on FP and FN rates generated from the ROC curves. Decision thresholds were selected at two distances in the probability distribution in which the

FP and FN rates were 0.5%, in order to minimize the number of FPs and FNs in the final model. A distance between two samples below 1.17 CAN distance indicated a linked association with a 99.5% confidence corresponding to a 0.5% FP rate, whereas a Canberra distance above 2.09 indicated an unlinked association with 99.5% confidence corresponding to a 0.5% FN rate. Between these two values lies the 'inconclusive' region, where linked and unlinked distributions significantly overlapped. A likely reason is the selection of the unlinked distances. The only criteria were that the samples should originate from different police cases. Amongst those, there are likely hitherto unknown pairs from the same batch, thus giving rise to linked distances. No attempts to remove those were done as removal of too much data may have resulted in overfitting the model, thereby reducing the suitability on unseen data.

To cross validate the method, a hierarchical clustered dendrogram and heatmap were calculated from 50 randomly selected cocaine samples from the entire group of samples (Figure 6). Similarities and differences in profile can be observed amongst the random samples, with many of the unrelated samples forming potentially linked pairs of clusters of observations. With comparison between individual samples on a case-by-case basis, these links may have remained unnoticed. Therefore, discovery of unknown links through retrospective comparison highlights a huge benefit of the developed methodology, as it can allow law enforcement to easily recognize links between new samples and old samples. A complete dendrogram and heatmap of all 483 cocaine samples are shown in Figure S3, displaying even more groups and links. However, due to the previously described limitations regarding selection of linked/unlinked populations, usage of additional validated methods is encouraged in order to confirm links, especially for comparisons producing distances in the inconclusive region.



**FIGURE 5** Boxplot (top) and histogram (bottom) of the standardized and normalized pretreated Canberra distances for the linked and unlinked pairwise groups



**FIGURE 6** Hierarchical clustered heatmap of 50 random samples from the unlinked population presenting a Canberra distance under the linked cut-off. Rows are the 12 target compounds from the final profiling model grouped by cluster analysis using complete linkage clustering and Canberra distance. Columns are the individual injections clustered using complete linkage clustering and cosine similarity. The colour scale of the heat map is standardized, and the normalized peak area is scaled from 0 to 1, with a dark colour representing a higher peak area and a light colour representing a lower peak area. A full heatmap with all 483 cocaine-positive samples can be found in Figure S1

Python-based scripts for data analysis and model development formed the core of the article. In this study, machine learning concepts such as unsupervised clustering algorithms were utilized; however, model development was mainly inspired by strategies adopted by others. Recent research has successfully produced a model for automatic cocaine classification using data-driven

approaches based on machine learning classification tools.<sup>39</sup> This demonstrates the possibility for further innovation within impurity profiling research by incorporating machine learning tools in the model development and comparison processes. Further research should therefore aim to explore the wider possibilities of these tools and aim to incorporate them within the research.

## 4 | CONCLUSION

In this study, a model was developed for the impurity profiling of cocaine based on retrospective UHPLC-TOF-MS data. A nontargeted marker discovery approach was used to discover 12 cocaine marker compounds that were used to develop the profiling model. Combinations of pretreatment and distance measures were evaluated based on ROC curves in order to select the most discriminatory statistical treatment for the dataset. Standardization with normalization transformation followed by Canberra distance was selected in the final profiling model, and a successful classification was achieved, with an AUC of 0.988. Linked and unlinked thresholds were selected using 99.5% confidence, resulting in linked and unlinked thresholds at 0.5% FP and FN rates, respectively. The model was developed with a novel approach to impurity profiling, utilizing chemometric-based tools in order to create an easily reproducible workflow. It allowed for development of a comparative profiling model without prior knowledge of links between samples, extensive sample preparation or prior knowledge of the target markers. The method relied solely on data retrospectively collected during routine drug screening, thereby requiring only reanalysis of a small number of samples evaluation of markers. Furthermore, the developed workflow was not specific to cocaine or UHPLC-TOF-MS data and therefore has potential applications with other seized drugs or with data collected from other analytical techniques. The present study demonstrates the possibility for further advances in the well-established field of impurity profiling analysis, to provide further intelligence support for law enforcement to aid in combatting drug manufacture and trafficking.

### ACKNOWLEDGEMENT

This research did not receive any specific grant from funding agencies in the public, commercial or not-for-profit sectors.

### CONFLICT OF INTEREST

The authors declare they have no conflicts of interest.

### DATA AVAILABILITY STATEMENT

Research data are not shared. The data are not publicly available due to privacy or ethical restrictions.

### ORCID

Daniel Carby-Robinson  <https://orcid.org/0000-0002-8912-8830>

Petur Weihe Dalsgaard  <https://orcid.org/0000-0002-8967-7477>

Christian Brinch Mollerup  <https://orcid.org/0000-0001-8187-6474>

### REFERENCES

1. Popovic A, Morelato M, Roux C, Beavis A. Review of the most common chemometric techniques in illicit drug profiling. *Forensic Sci Int*. 2019;302:109911. <https://doi.org/10.1016/j.forsciint.2019.109911>
2. United Nations Office on Drugs and Crime. Methods for impurity profiling of heroin and cocaine; 2005. Accessed May 3, 2021.
3. United Nations Office for Drug Control and Crime Prevention. Drug characterization/impurity profiling: background and concepts; 2001. Accessed May 3, 2021.
4. Broséus J. Chemical profiling: a tool to decipher the structure and organisation of illicit drug markets: an 8-year study in Western Switzerland. *Forensic Sci Int* Published online. 2016;266:18-28.
5. Collins M, Huttunen J, Evans I, Robertson J. Illicit drug profiling: the Australian experience. *Aust J Forensic Sci*. 2007;39(1):25-32. <https://doi.org/10.1080/00450610701324924>
6. Locicero S, Hayoz P, Esseiva P, Dujourdy L, Besacier F, Margot P. Cocaine profiling for strategic intelligence purposes, a cross-border project between France and Switzerland. *Forensic Sci Int*. 2007;167(2-3):220-228. <https://doi.org/10.1016/j.forsciint.2006.06.052>
7. Broséus J, Huhtala S, Esseiva P. First systematic chemical profiling of cocaine police seizures in Finland in the framework of an intelligence-led approach. *Forensic Sci Int*. 2015;251:87-94. <https://doi.org/10.1016/j.forsciint.2015.03.026>
8. European Monitoring Centre for Drugs and Addiction. European Drug Report 2020: Trends and Developments; 2020. [https://www.emcdda.europa.eu/system/files/publications/13236/TDAT20001ENN\\_web.pdf](https://www.emcdda.europa.eu/system/files/publications/13236/TDAT20001ENN_web.pdf). Accessed May 3, 2021.
9. Mallette JR, Casale JF, Jones LM, Morello DR. The isotopic fractionation of carbon, nitrogen, hydrogen, and oxygen during illicit production of cocaine base in South America. *Forensic Sci Int*. 2017;270:255-260. <https://doi.org/10.1016/j.forsciint.2016.10.016>
10. Casale JF, Klein RFX. Illicit production of cocaine. *Forensic Sci Rev*. 1993;5(2):95-107.
11. United Nations Division of Narcotic Drugs. Recommended Methods for Testing Cocaine: Manual for Use by National Narcotics Laboratories; 1986. Accessed May 3, 2021.
12. Casale JF, Waggoner RW. A chromatographic impurity signature profile analysis for cocaine using capillary gas chromatography. *J Forensic Sci*. 1991;36(5):13154J. <https://doi.org/10.1520/JFS13154J>
13. Mollerup CB, Dalsgaard PW, Mardal M, Linnet K. Targeted and nontargeted drug screening in whole blood by UHPLC-TOF-MS with data-independent acquisition: Targeted and non-targeted drug screening in whole blood by UHPLC-TOF-MS. *Drug Test Anal*. 2017;9(7):1052-1061. <https://doi.org/10.1002/dta.2120>
14. Pedersen AJ, Dalsgaard PW, Rode AJ, et al. Screening for illicit and medicinal drugs in whole blood using fully automated SPE and ultra-high-performance liquid chromatography with TOF-MS with data-independent acquisition: liquid chromatography. *J Sep Sci*. 2013;36(13):2081-2089. <https://doi.org/10.1002/jssc.201200921>
15. Vincenti F, Montesano C, di Ottavio F, et al. Molecular networking: a useful tool for the identification of new psychoactive substances in seizures by LC-HRMS. *Front Chem*. 2020;8:572952. <https://doi.org/10.3389/fchem.2020.572952>
16. Odoardi S, Valentini V, de Giovanni N, Pascali VL, Strano-Rossi S. High-throughput screening for drugs of abuse and pharmaceutical drugs in hair by liquid-chromatography-high resolution mass spectrometry (LC-HRMS). *Microchem J*. 2017;133:302-310. <https://doi.org/10.1016/j.microc.2017.03.050>
17. Seither JZ, Hindle R, Arroyo-Mora LE, DeCaprio AP. Systematic analysis of novel psychoactive substances. I. Development of a compound database and HRMS spectral library. *Forensic Chem*. 2018;9:12-20. <https://doi.org/10.1016/j.forc.2018.03.003>
18. Mollerup CB, Rasmussen BS, Johansen SS, Mardal M, Linnet K, Dalsgaard PW. Retrospective analysis for valproate screening targets with liquid chromatography-high resolution mass spectrometry with positive electrospray ionization: an omics-based approach. *Drug Test Anal*. 2019;11(5):730-738. <https://doi.org/10.1002/dta.2543>
19. Katajamaa M, Oresic M. Data processing for mass spectrometry-based metabolomics. *J Chromatogr A*. 1158:318-328.
20. McKinney W. Data structures for statistical computing in python. 2010:56-61. <https://doi.org/10.25080/Majora-92bf1922-00a>
21. Harris CR, Millman KJ, van der Walt SJ, et al. Array programming with NumPy. *Nature*. 2020;585(7825):357-362. <https://doi.org/10.1038/s41586-020-2649-2>

22. SciPy 1.0 Contributors, Virtanen P, Gommers R, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat Methods*. 2020;17(3):261-272. <https://doi.org/10.1038/s41592-019-0686-2>
23. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine Learning in Python. *Mach Learn PYTHON*. 2011;12:2825-2830.
24. Tukey JW. *Exploratory Data Analysis*. 7th ed. Reading, MA: Addison-Wesley Pub Co; 1997.
25. Hunter JD. Matplotlib: a 2D graphics environment. *Comput Sci Eng*. 2007;9(3):90-95. <https://doi.org/10.1109/MCSE.2007.55>
26. Chicco D, Jurman G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics*. 2020;21(1) 6. <https://doi.org/10.1186/s12864-019-6413-7>
27. Bovens M, Ahrens B, Alberink I, Nordgaard A, Salonen T, Huhtala S. Chemometrics in forensic chemistry—Part I: implications to the forensic workflow. *Forensic Sci Int*. 2019;301:82-90. <https://doi.org/10.1016/j.forsciint.2019.05.030>
28. Salonen T, Ahrens B, Bovens M, et al. Chemometrics in forensic chemistry—Part II: standardized applications—three examples involving illicit drugs. *Forensic Sci Int*. 2020;307:110138. <https://doi.org/10.1016/j.forsciint.2019.110138>
29. Esseiva P, Gaste L, Alvarez D, Anglada F. Illicit drug profiling, reflection on statistical comparisons. *Forensic Sci Int*. 2011;207(1-3):27-34. <https://doi.org/10.1016/j.forsciint.2010.08.015>
30. Bradley AP. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognit*. 1997;30(7):1145-1159. [https://doi.org/10.1016/S0031-3203\(96\)00142-2](https://doi.org/10.1016/S0031-3203(96)00142-2)
31. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*. 1982;143(1):29-36. <https://doi.org/10.1148/radiology.143.1.7063747>
32. Vinkovic K, Galic N, Schmid MG. Micro-HPLC-UV analysis of cocaine and its adulterants in illicit cocaine samples seized by Austrian police from 2012 to 2017. *J Liq Chromatogr Relat Technol*. 2018;41(1):6-13. <https://doi.org/10.1080/10826076.2017.1409237>
33. Broséus J, Gentile N, Esseiva P. The cutting of cocaine and heroin: a critical review. *Forensic Sci Int*. 2016;262:73-83. <https://doi.org/10.1016/j.forsciint.2016.02.033>
34. Fiorentin TR, Krotulski AJ, Martin DM. Detection of cutting agents in drug-positive seized exhibits within the United States. *J Forensic Sci*. 2019;64(3):888-896. <https://doi.org/10.1111/1556-4029.13968>
35. Stride Nielsen L, Lindholm C, Villesen P. Cocaine classification using alkaloid and residual solvent profiling. *Forensic Sci Int*. 2016;269:42-49. <https://doi.org/10.1016/j.forsciint.2016.11.007>
36. Liu C, Hua Z, Meng X. Applicability of ultra-high performance liquid chromatography-quadrupole-time of flight mass spectrometry for cocaine profiling: profiling of cocaine by UHPLC-QTOF-MS. *Drug Test Anal*. 2017;9(8):1152-1161. <https://doi.org/10.1002/dta.2132>
37. Locicero S, Esseiva P, Hayoz P, Dujourdy L, Besacrier F, Margot P. Cocaine profiling for strategic intelligence, a cross-border project between France and Switzerland. *Forensic Sci Int*. 2008;177(2-3):199-206. <https://doi.org/10.1016/j.forsciint.2007.12.008>
38. Dujourdy L, Besacrier F. Headspace profiling of cocaine samples for intelligence purposes. *Forensic Sci Int*. 2008;179(2-3):111-122. <https://doi.org/10.1016/j.forsciint.2008.04.024>
39. Cascini F. A data-driven methodology to discover similarities between cocaine samples. *Sci Rep* Published online. 2020;12.
40. Kruev A, Kaupmees K. Adduct formation in ESI/MS by mobile phase additives. *J Am Soc Mass Spectrom*. 2017;28(5):887-894. <https://doi.org/10.1007/s13361-017-1626-y>
41. Mallette JR, Casale JF. Rapid determination of the isomeric truxillines in illicit cocaine via capillary gas chromatography/flame ionization detection and their use and implication in the determination of cocaine origin and trafficking routes. *J Chromatogr a*. 2014;1364:234-240. <https://doi.org/10.1016/j.chroma.2014.08.072>
42. Cui X, Wang R, Lian R, Liang C, Chen G, Zhang Y. Correlation analysis between cocaine samples seized in China by the rapid detection of organic impurities using direct analysis in real time coupled with high-resolution mass spectrometry. *Int J Mass Spectrom*. 2019;444:116188. <https://doi.org/10.1016/j.ijms.2019.116188>
43. Liu C, Hua Z, Bai Y. Classification of illicit heroin by UPLC-Q-TOF analysis of acidic and neutral manufacturing impurities. *Forensic Sci Int*. 2015;257:196-202. <https://doi.org/10.1016/j.forsciint.2015.08.009>
44. Morelato M, Beavis A, Tahtouh M, Ribaux O, Kirkbride P, Roux C. The use of organic and inorganic impurities found in MDMA police seizures in a drug intelligence perspective. *Sci Justice*. 2014;54(1):32-41. <https://doi.org/10.1016/j.scijus.2013.08.006>
45. Kuwayama K, Tsujikawa K, Miyaguchi H, et al. Identification of impurities and the statistical classification of methamphetamine using headspace solid phase microextraction and gas chromatography-mass spectrometry. *Forensic Sci Int*. 2006;160(1):44-52. <https://doi.org/10.1016/j.forsciint.2005.08.013>

#### SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

**How to cite this article:** Carby-Robinson D, Dalsgaard PW, Mollerup CB, Linnet K, Rasmussen BS. Cocaine profiling method retrospectively developed with nontargeted discovery of markers using liquid chromatography with time-of-flight mass spectrometry data. *Drug Test Anal*. 2022;14(3):462-473. <https://doi.org/10.1002/dta.3130>