Advances in the Chemometric Analysis of Multiway Chromatographic Data to Improve Discovery and Identification

Caitlin N. Cain

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2024

Reading Committee:

Robert Synovec, Chair

Nicholas Riley

Bo Zhang

Program Authorized to Offer Degree:

Chemistry

University of Washington

Abstract

Advances in the Chemometric Analysis of Multiway Chromatographic Data to Improve Discovery and Identification

Caitlin N. Cain

Chair of the Supervisory Committee:

Robert Synovec

Chemistry

Both one-dimensional gas chromatography (1D-GC) and comprehensive two-dimensional gas chromatography (GC×GC) are used in various applications because of their ability to discover and identify pure chemical species in volatile and semi-volatile mixtures. However, the information-rich data sets produced by these instruments, especially when they are coupled to mass spectrometry (MS), are often too large and complex to manually interpret. Therefore, the application of advanced data analysis methods, referred to as chemometrics, are necessary to efficiently analyze and extract meaningful chemical information from these instrumental platforms. This dissertation presents the development and application of several novel chemometric approaches that improve the discovery and identification of key chemical species in both 1D-GC-MS and GC×GC-MS data sets. First, this dissertation describes the application of

Fisher ratio (F-ratio) analysis to create a chemical fingerprint of potato taste defect in roasted coffee beans and thermal stress in kerosene-based rocket fuels. As a supervised chemometric technique, F-ratio analysis utilizes prior knowledge of sample class membership to discover statistically significant concentration differences in chromatographic data sets. However, knowledge about the samples or experimental design may not be available during analysis. To address this situation, this dissertation describes the development of two unsupervised data analysis approaches. For large chromatographic data sets, variance ranking analysis was created to discover analytes exhibiting a high signal variance across the samples. Application of variance ranking analysis, along with principal components analysis and *k*-means clustering, to multiple metabolomic data sets uncovered hidden chemical patterns and sample groupings. Variance ranking analysis was also demonstrated to be an effective data reduction technique for developing accurate physicochemical models of aerospace fuels with partial least squares regression. For studies that may be limited in the number of samples and/or chromatographic replicates, a pairwise analysis method known as 1v1 analysis was developed to find chemical differences between two chromatograms. This method can also extract a purified mass spectrum to improve compound identifications for analytes at low chromatographic resolutions and/or with high signal interferences. The performance of both unsupervised analyses was shown to be comparable to F-ratio analysis. Finally, this dissertation also advances the capacity to reliably discover and identify analytes using a single chromatogram. The generation of an enhanced total ion current chromatogram (TIC) is introduced to improve visualization of analytical signals previously obscured by the background noise. The enhanced TIC algorithm improves the detection of analytical signals by denoising the mass spectral dimension. Concurrently, an intra-mass channel (*m/z*) comparison method, termed mzCompare, is developed to improve the

identification of unresolved chemical species. This approach generates pure analyte profiles for unresolved chemical species by discovering *m/z* with similar retention times and peak shapes. These purified profiles are then used as a constraint in a chemometric decomposition model to mathematically resolve the overlapped species and achieve accurate compound identifications.

Table of Contents

List of Figures

List of Tables

Acknowledgements

This dissertation would not have been possible without the mentorship and guidance I have received during my academic career. First and foremost, I would like to thank my Ph.D. advisor, Dr. Robert Synovec, for his support and encouragement throughout my projects here at the University of Washington (UW). He has continued to push and challenge me to explore new areas of chromatography, chemometrics, and analytical chemistry that I never would have imagined. The work presented herein is a culmination of both the research and leadership skills that I have developed while in his group. I would also like to thank everyone that has served on either my second-year, general, and final examination committees for their input and advice on my Ph.D. studies. Lastly, I must acknowledge my undergraduate research mentors at Virginia Commonwealth University, Drs. Maryanne Collinson and Sarah Rutan. They ignited my initial interest in separation science by accepting me into their laboratories during my freshman year and treating my independent research projects no differently than their graduate students. Where I am today is largely due to their encouragement.

Throughout my past five years in the Synovec group, I have had the opportunity to work alongside some of the brightest people and I am thankful for all their collaboration and feedback on my research. Drs. Paige Sudol, Sonia Schöneich, Grant Ochoa, and Timothy Trinklein – Thank you for also becoming my best friends during this process. I appreciate all the time we have spent together goofing around in the office and grabbing coffee when we were all feeling the afternoon hit. I am also thankful for all the time we spent together outside of the office, whether it was at Shultzy's, different breweries in Seattle, Pure Barre, or the climbing gym. For the past year, I have had the opportunity to share an office with one of the newest members of

the Synovec group - Austin Dobrecevich (Augie D). Thank you for taking a genuine interest in my work, listening to me vent about formatting this dissertation, and sharing with me all the West Seattle updates. I look forward to seeing everything you accomplish in the future. I am also grateful to have met several amazing people in the Atmospheric Sciences department – Ursula Jongebloed, Vince Cooper, Phil Rund, Adam Sokol, and Kaitlyn Confer – during my time here at UW. I can confidently say that our friendship has taught me more than I ever wanted to know about aerosols, climate dynamics, and cloud-climate interactions, and I am going to miss our late night summer hangouts either in someone's backyard or at Golden Gardens.

Finally, a brief paragraph does not even begin to describe how thankful I am to my family for their support and encouragement during my Ph.D. I will forever be grateful to my parents, Kim and Rod, for dedicating everything to my education while growing up. I know that I made you all drive across Virginia to take me to every nerd camp and science fair, but those experiences are why I am writing this dissertation today. Thank you for instilling in me the drive, work ethic, and stubbornness that made this Ph.D. possible. Thank you for the endless supply of puppy pictures and out-of-context GIFs when I needed to smile or laugh during a stressful day. Speaking of, I got to give a huge shout out to Zuri (rest in peace), Bindi, and Sydney for being all-around good puppies. I would also like to thank my future in-laws, Janet and Mark, for their encouragement and excitement whenever they heard about my successes. Last but certainly not least, I must give the biggest thank you in the world to my fiancé, Joseph Robinson. Thank you for helping me debug my Matlab code and rephrase things when I cannot get my thoughts right on the page. Thank you for being my biggest cheerleader, believing in me, and pushing me to dream bigger. Thank you for always reminding me that "when the work interferes with the fun, the work must go undone". I love you and I look forward to our next chapter.

Dedication

To my parents and fiancé for their love and support.

Chapter 1: Introduction to Chromatographic Separations and Chemometric Analysis

1.1. Introduction to Gas Chromatography

The development and application of new experimental methods, instrumental techniques, and computational strategies to identify and quantify compounds in complex matrices has been the long-standing mission for analytical chemistry. Selecting the appropriate sample preparation, instrumental, and data analysis methods is primarily driven by the research question and/or goal. The analytical objective for a research study can be broken down into two categories, targeted or discovery-based (i.e., non-targeted). Targeted analyses aim to identify and quantify the presence of a specific (or subset of) compound(s) in a sample. Conversely, discovery-based analyses aim to develop a comprehensive profile of all the compounds present in a sample. These analytical objectives not only drive the selection of appropriate instrumental and computational techniques, but they can also further innovations in the field of analytical chemistry.

Chromatography has been a fundamental instrumental technique for resolving compounds (i.e., analytes) in a sample, with the goal of achieving accurate identification and quantitation. Chromatographic separations are based on the partitioning of analytes between the mobile and stationary phases. The stationary phase can be defined as a solid or liquid phase coated on or inside an immobile support (e.g., a column). Meanwhile, the mobile phase is typically a liquid or gas that moves a mixture of analytes through the stationary phase. The

---

Parts of this chapter are reproduced from:

C.N. Cain, T.J. Trinklein, S. Schöneich, G.S. Ochoa, S.C. Rutan, R.E. Synovec, Comprehensive Two-Dimensional Chromatography with Chemometric Data Analysis, in: N. Grinberg, P.W. Carr (Eds.), Adv. Chromatogr., CRC Press, Boca Raton, 2022: pp. 145–191.

C.N. Cain,* S. Schöneich,* R.E. Synovec, Recent Advances in Comparative Analysis for Comprehensive Two-Dimensional Gas Chromatography–Mass Spectrometry Data, in: A.C. Olivieri, G.M. Escandar, H.C. Goicoechea, A. Munoz de la Pena (Eds.), Fundam. Appl. Multiw. Data Anal., Elsevier, Amsterdam, Netherlands, 2024: pp. 465–515. * These authors contributed equally.

partitioning of analytes between the mobile and stationary phases can be described as an equilibrium and is mathematically defined by the distribution coefficient (K_D),

$$K_D = \frac{[analyte]_{SP}}{[analyte]_{MP}} \quad (1.1)$$

where $[analyte]_{SP}$ and $[analyte]_{MP}$ represent concentrations of an analyte in the stationary phase and mobile phase, respectively. Hence, analytes with a larger K_D will have a stronger affinity for the stationary phase while analytes with a smaller K_D will have a stronger affinity for the mobile phase.

The retention factor (k') is a measure of the time that an analyte spends in the stationary phase relative to the time it spends in the mobile. The k' of an analyte can be expressed as

$$k' = \frac{t_R - t_0}{t_0} \quad (1.2)$$

where t_R describes the time between the sample injection into the chromatographic system and the elution of the analyte from the stationary phase (i.e., the retention time) and t_0 describes the time required for a completely unretained analyte to travel through the stationary phase (i.e., the dead time). The retention factor is related to the K_D by the following relationship,

$$k' = K_D \times \frac{V_{SP}}{V_{MP}} \quad (1.3)$$

where V_{SP} and V_{MP} are the stationary phase and mobile phase volumes, respectively. Thus, the mobile and stationary phases must be carefully optimized to ensure that each analyte will have a different K_D and produce a different k' during a chromatographic separation.

While many types of separations exist in the literature, this chapter will focus on the use of gas chromatography (GC). In GC, volatile and/or semi-volatile analytes partition between an inert, carrier gas (the mobile phase) and a thin polymer layer coated onto a fused silica capillary column (the stationary phase), which is housed inside a thermostatted oven. Typical carrier gases for GC are hydrogen or helium. Depending on the application, the thin polymer layer of the

stationary phase is either a substituted, cross-linked polydimethylsilicone (PDMS) or polyethylene glycol (PEG) backbone. The typical GC column can be 15 – 60 m long, with a stationary phase inner diameter of 100 – 320 μm, and film thickness of 0.1 – 0.5 μm. Since the carrier gas is inert, analyte partitioning in GC is only dependent on stationary phase affinity and volatility (i.e., K_D is only dependent on $[\text{analyte}]_{SP}$). For a homologous series, compounds will elute from the GC column in order of their boiling points. However, for mixtures containing analytes with different functional groups, the elution order will be dependent on the interrelationship of stationary phase affinity and boiling point. Ultimately, the retention of an analyte can be adjusted by tuning the GC column parameters, stationary phase composition, or oven temperature.

Following a GC separation, the separated analytes reach a detector, which converts the chemical concentration into an electronic signal measurement. Typically, signal measurements for separated analytes are recorded as Gaussian-like peaks. Figure 1.1 illustrates the Gaussian-like profile of a chromatographic peak, where the t_R of analyte is measured at the maximum peak height and the peak width is measured at either baseline (W_b) or half-height ($W_{1/2}$). Detectors for GC can be either univariate, which records a single signal during the separation, or multivariate, which records multiple signals during the separation. Common univariate detectors for GC include flame ionization detector (FID) [1–3] or the electron capture detector (ECD) [4,5]. However, analyte identification with univariate detectors can be challenging, relying solely on t_R matching of sample peaks to a known standard. MS is especially useful for identification of analytes in non-targeted studies and for obtaining pure mass channels (m/z) that can be used for accurate analyte quantification. Common MS detectors used for GC instrumental platforms are the time-of-flight mass spectrometer (TOFMS) [6–8] and quadrupole mass spectrometer (qMS)

[9–11]. The common nominal mass TOFMS detectors can achieve high collection rates (up to 500 Hz) with high sensitivity and selectivity, which is especially advantageous for collecting enough data points across narrow GC peak widths. Alternatively, while the qMS operates at lower data collection frequencies (up to 50 Hz for a fast scanning qMS), it is less expensive to maintain while still being useful for a large range of applications. Note, unlike the TOFMS, quadrupole analyzers do not scan all *m/z* simultaneously. Instead, qMS detectors scan at a rate dependent upon the scan cycle time, which includes the dwell time, or how long it scans each *m/z*, and the interscan delay, the time between successive scans of the *m/z* range. Therefore, *m/z* which belong to same analyte reach their respective maxima at different times and will need to be corrected during any data preprocessing steps (discussed later in *1.3.1.2. Chromatographic data* preprocessing).

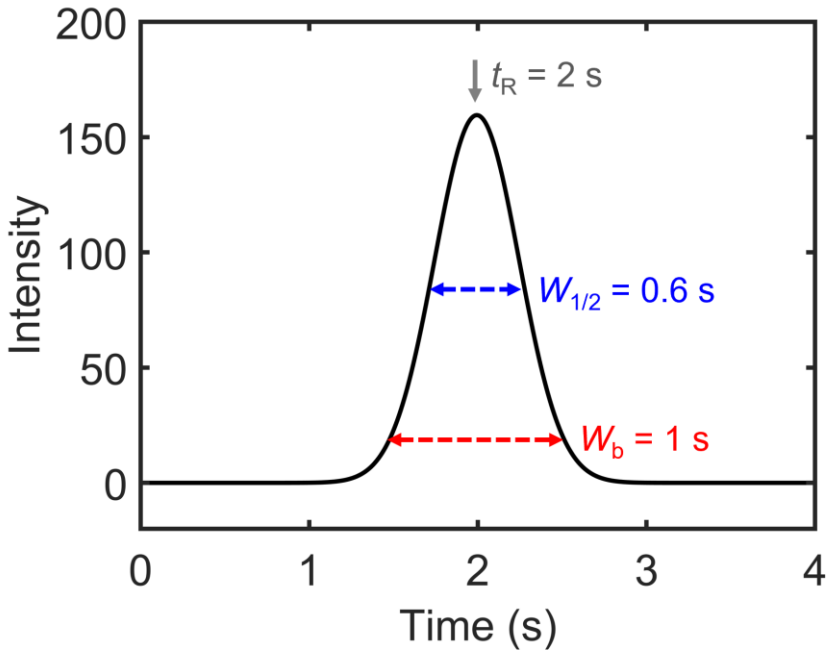


Figure 1.1. Illustration of retention time (t_R) and peak width (at baseline, W_b, and at half-height, $W_{1/2}$) measurements on a simulated chromatographic peak.

An important figure of merit in chromatography is resolution (R_s), which describes the degree of overlap between two adjacent peaks. Mathematically, R_s is defined as

$$R_s = \frac{2(t_{R,2}-t_{R,1})}{W_{b,2}+W_{b,1}} = \frac{1.18(t_{R,2}-t_{R,1})}{W_{\frac{1}{2},2}+W_{\frac{1}{2},1}} \quad (1.4)$$

Figure 1.2 illustrates the R_s between two analytes (red curve and blue curve), along with what the signal that the detector measures (black trace), at four different values: (A) 1.5, (B) 1.0, (C) 0.5, and (D) 0.2. A R_s of 1.5 (Figure 1.2A) is required to achieve baseline resolution between two analytes, which ensures accurate analyte identification and quantitation. However, an R_s = 1.5 is difficult to achieve in real, complex mixtures. Therefore, unit resolution (R_s = 1; Figure 1.2B) is typically accepted for identification and quantitation (only using peak heights) since the mutual overlap between these two peaks equals 2.3 %. However, at R_s < 1 (Figure 1.2C-D), the detector is unable to distinguish between the two analytes and thus, only one peak appears in the chromatogram (black trace). The use of a multivariate detector like MS and/or chemometric decomposition software (discussed later in *1.3.2. Targeted chemometric methods*) would be required to determine that this peak contained two analytes.

The goal for any chromatographic separation is to obtain the maximal R_s between analytes in the least overall separation time. However, achieving this goal in isothermal GC separations is a challenge. Low oven temperatures will not only ensure that all analytes have a high R_s, but it will also increase the overall separation time and analytes with a high *k*' will have wider W_b, decreasing their overall signal-to-noise ratio (*S/N*). Meanwhile, higher oven temperatures will decrease the analysis time and W_b, but it will also decrease R_s. This challenge is referred to as the general elution problem. To overcome this the general elution problem, temperature programming is employed. With temperature programming, the oven temperature increases over the course of the separation, to maximize R_s while minimizing W_b.

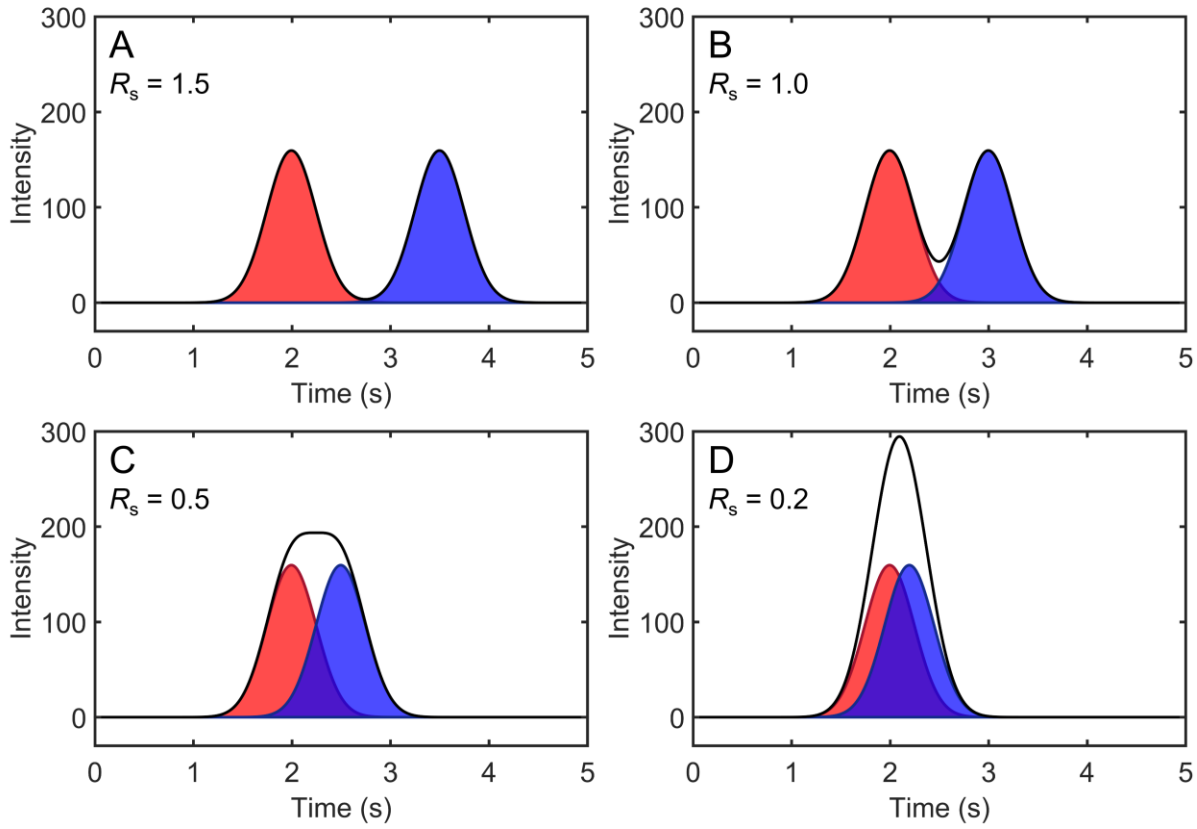


Figure 1.2. Illustration of chromatographic resolution (R_s) for two analytes (red and blue) at different values: (A) 1.5, (B) 1.0, (C) 0.5, and (D) 0.2. The black trace is the signal that a detector would measure as these analytes elute from the GC column.

Peak capacity (n_c) is another key figure of merit for assessing the overall performance of a chromatographic separation and is defined as

$$n_c = \frac{t_{sep}}{W_b R_s} \tag{1.5}$$

where t_{sep} defines the separation time window. At unit resolution (R_s = 1), this metric equals the number of peaks that can be separated in a chromatographic run. Thus, producing narrow W_b through temperature programming and optimizing the GC column dimensions will increase n_c along with R_s. However, as research has shifted from targeted to non-targeted studies, the number of analytes in these complex samples has exceeded the resolving power of one-dimensional GC (1D-GC). In 1983, Davis and Giddings introduced statistical overlap theory

(SOT) to highlight the insufficient peak capacity provided by 1D separations for multicomponent mixtures [12]. Specifically, SOT found that when the number of analytes in a sample equals the peak capacity (i.e., a saturation factor of 1), the maximum number of resolvable peaks is only 37 % of the peak capacity [12]. Note, the term peak in this context refers to a distinct concentration pulse that can either be due to a single analyte or several analytes. For resolvable, single-analyte peaks, SOT suggests that the maximum number observable at a saturation factor of 1 is limited to 18 % of the peak capacity [12]. For example, the 1D-GC chromatogram of a petroleum-based fuel, which is comprised of comprising hundreds to thousands of analytes with similar boiling points and chemical substitutions, often exhibits a large "hump" in the baseline. This is termed an unresolved complex mixture, which hinders both compound identification and quantitation. Fortunately, the use of two (or more) separations in sequence can overcome these statistical limitations and analytical challenges.

1.2. Introduction to Comprehensive Two-Dimensional Gas Chromatography

Multidimensional gas chromatography (MDGC) utilizes two GC separations to improve R_s. Many early MDGC adopters reported the use of heart-cutting, where a subset of the analytes from the 1D separation are subjected to secondary separation [13]. Heart-cutting is particularly useful for targeting a known set of analytes of interest and discarding the remaining information about the samples. However, for non-targeted studies, comprehensive two-dimensional (2D) gas chromatography (GC×GC) is especially useful for the full characterization of samples without prior knowledge of analytes of interest. Liu and Phillips first introduced GC×GC in 1991 [14]. A GC×GC separation is achieved by connecting two columns with sufficiently orthogonal stationary phase selectivity in series and interfaced with a secondary injector, termed the modulator. Typically, the first dimension (^{1}D) column is the typical length of 20 – 30 m like in

1D-GC while the second dimension (^{2}D) column is kept shorter at 1 – 5 m long. The modulator periodically traps or focuses effluent from the ^{1}D column effluent and reinjects it onto the head of the second dimension ^{2}D column in a sharp pulse. For a separation to be sufficiently comprehensive, each ^{1}D peak must be sampled a minimum of 2 – 4 times [15,16]. The time between sampling events is denoted by the modulation period, P_M. The typical P_M (i.e., the ^{2}D separation time) ranges from 1 – 6 s. The use of a short P_M ensures that the resolution of the separation that was achieved on ^{1}D is not seriously degraded due to undersampling.

Modulators for GC×GC can be broadly classified into two groups: thermal modulators and flow modulators, also known as valve-based or pneumatic modulators. Thermal modulators use cold and hot pulses to trap and re-inject analytes onto the ^{2}D column with increased sensitivity due to the thermal focusing of analytes in the cold zone and total transfer of effluent from the ^{1}D column to the ^{2}D column. Initially, cryogenic thermal modulation was used to achieve the short cold pulses; however, a disadvantage of this modulation technique was the cost of using liquid cryogen. Hence, cryogen-free thermal modulation was introduced to mitigate this issue [17–22]. Commercial GC×GC-MS instruments are often equipped with thermal modulators as they are compatible with the MS detector pumping capacity. Further, thermal modulation provides reproducible peak shapes and excellent sensitivity improvement relative to 1D-GC due to the focusing of analytes from the ^{1}D column. The resulting data produced generally does not require extensive data processing prior to chemometric analysis. However, along with a higher cost of operation, thermal modulators are not able to modulate light, highly volatile compounds ($< C_5$) unlike flow modulators.

Flow modulators, typically flow diversion or differential flow, use an auxiliary flow of carrier gas and re-direct, collect, or temporarily stop flow from the ^{1}D column onto the ^{2}D

column [23–32]. These modulators are less expensive to implement than thermal modulators; however, not all flow modulators can achieve a 100 % duty cycle and the high ^{2}D flow rates (^2F) required for some differential flow modulators to flush sample out of the sample loop onto the ^{2}D column are not compatible with the pumping capacity of MS detectors. To address the high ^2F, one solution is to split flow to multiple detectors, but splitting flow will cause a decrease sensitivity. Alternatively, significant progress has been made to ensure flow modulators can be operated at flow rates compatible with MS detectors [23,32–39]. Other concerns using flow modulation is the reproducibility of the peak shape for a given analyte [32,40] and under-sampling of the ^{1}D effluent on quantitative precision [41]. Both effects can negatively impact the data quality and hinder the performance of chemometric techniques on the data set. Therefore, it is important to select a proper flow ratio and modulation ratio to mitigate these concerns.

Regardless of the modulator selected for a GC×GC instrument, Figure 1.3 demonstrates the overall modulation process for two analytes (green and red curves) unresolved by 1D-GC. However, in GC×GC, this peak would be sampled multiple times by the modulator as it elutes from the ^{1}D column and reinjected onto the ^{2}D column. This is observed in the unfolded, GC×GC chromatogram that is measured by the detector, with the dashed lines corresponding to the P_M of 2 s. In this GC×GC setup, the green and red analyte are resolved from one another due to the complementary selectivity of the ^{2}D column. Additionally, it is important to note that both analytes also had their signals enhanced in the GC×GC separation compared to the 1D-GC separation due to the modulation process (Figure 1.3). The improved selectivity and increased sensitivity are key advantages of performing a GC×GC separation [42,43]. Ultimately, the resulting data array can be transformed into a three-dimensional (3D) chromatogram by cutting the data array at the time of each P_M sampling event and stacking those events alongside one

another (Figure 1.3). For improved visualization, the 3D waterfall plot in can be collapsed into a 2D contour plot (Figure 1.3). Here, both the t_R (^1t_R and ^2t_R) and W_b (^1W_b and ^2W_b) for each analyte on each separation dimension can easily be measured.

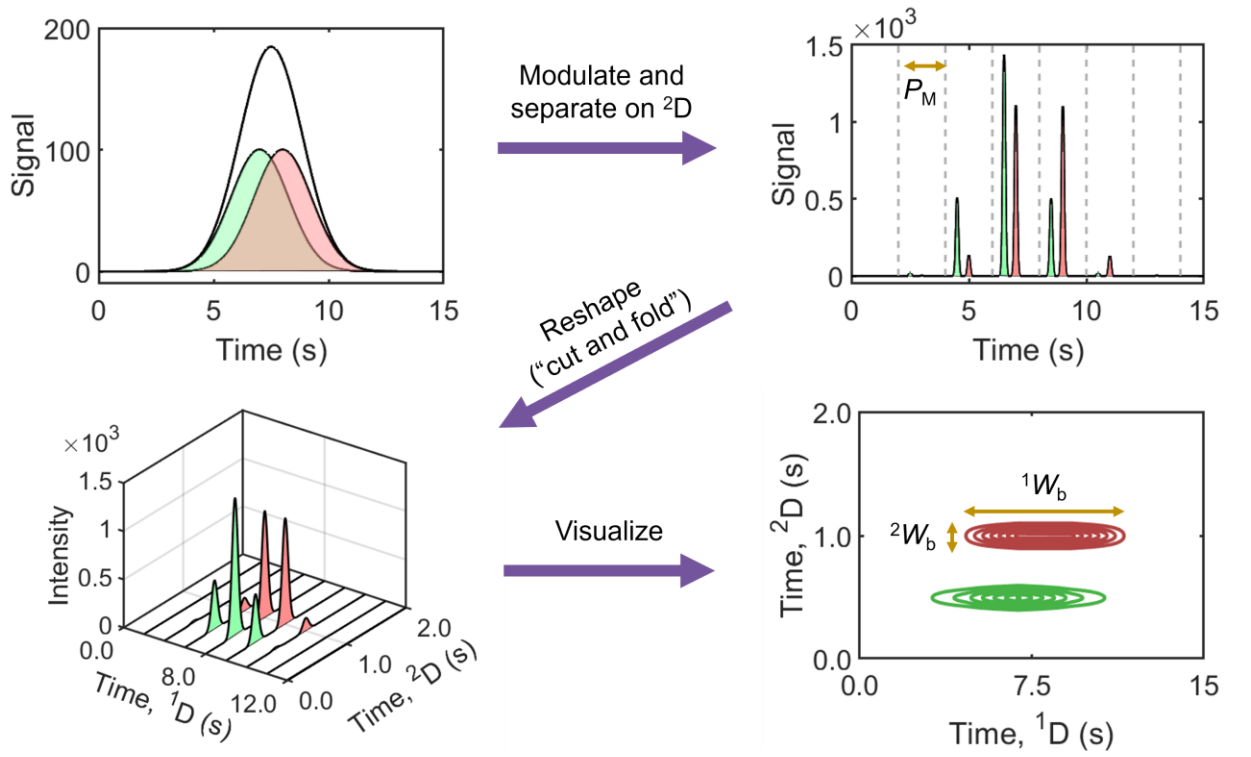


Figure 1.3. Illustration of the modulation process in GC×GC for two unresolved analytes (green and red). As the analytes elute from the ^{1}D column, they are sampled and reinjected onto the ^{2}D column by the modulator, operating at a user specified modulation period (P_M). The two analytes are resolved by the complementary ^{2}D column. The black trace demonstrates the signal that a detector would record after the ^{1}D and ^{2}D column. Each P_M can then be cut out from the data and stacked alongside each other to visualize the data as either a 3D waterfall plot or 2D contour plot.

As demonstrated in Figure 1.3, a comprehensive 2D separation like GC×GC has the capability to improve R_s for analytes that are highly overlapped on the ^{1}D. The R_s ($R_{s,2D}$) between two peaks in a comprehensive 2D separation is measured as the Euclidean norm between the R_s on the ^{1}D and ^{2}D (^1R_s and ^2R_s), which is defined as [15]

$$R_{s,2D} = \sqrt{{}^1R_s{}^2 + {}^2R_s{}^2} = \sqrt{\frac{2({}^1t_{R,2} - {}^1t_{R,1})}{{}^1W_{b,2} + {}^1W_{b,1}} + \frac{2({}^2t_{R,2} - {}^2t_{R,1})}{{}^2W_{b,2} + {}^2W_{b,1}}} \quad (1.6)$$

The ability of a comprehensive 2D separation to improve R_s between overlapped analytes is due to its enhanced n_c compared to a 1D separation. The ideal n_c for a comprehensive 2D separation ($n_{c,2D}$), defined at unit $R_{s,2D}$, is

$$n_{c,2D} = {}^1n_c \times {}^2n_c = \frac{{}^1t_{sep}}{{}^1W_b} \times \frac{P_M}{{}^2W_b} \quad (1.7)$$

where ${}^1n_c$ and ${}^2n_c$ describe the peak capacity on the ^{1}D and ^{2}D, respectively. As both Eqs. 1.6 and 1.7 show, the key to achieving high $R_{s,2D}$ and $n_{c,2D}$ relies upon generating narrow peak widths on both dimensions. For an ideal GC×GC separation, Klee *et al*. demonstrated a ~ 10-fold increase in n_c provided by the 2D separation compared to its 1D counterpart [44]. To achieve these near-ideal increases in n_c, the separations in each dimension should be complementary and sufficiently independent, which in some cases can provide compound group-type separations that aid in the interpretation of these chromatograms. Given the increased resolving power, the use of GC×GC has been growing for complex applications including fuel [45,46], food [47,48], cosmetic [49,50], forensic [51,52], metabolomic [8,53], and environmental samples [54,55].

1.3. Introduction to Chemometrics

Conventional data analysis approaches for 1D-GC or GC×GC involve an analyst manually identifying, quantifying, and determining the significance for every peak in the chromatogram in all samples. Due to the overall size and complexity of these data sets, this hands-on data analysis process can become quite onerous and time-consuming, especially for GC×GC data. Furthermore, conventional data analysis approaches are often not feasible for GC×GC separations since each compound results in 2 – 4 peaks in sequential ^{2}D chromatograms. Therefore, historical practices relied on targeting either a specific group of analytes (or compound classes) [56,57]. In this targeted approach, the experimental conditions are tailored towards the analytes of interest and the remainder of the sample is not considered. While targeted

approaches may achieve the initial goal of the analyst, a large amount of information about the samples remains unknown and unused. Discovery-based studies, on the other hand, can be beneficial in comprehensively analyzing the chemical information embedded in a 1D-GC or GC×GC data set. To achieve this goal, these studies depend on the use of chemometrics, which can effectively and efficiently analyze the data with less analyst intervention and in a shorter amount of time compared to manual approaches.

Here, chemometrics refers to the use of linear algebra and statistical methods to extract meaningful chemical information from analytical data sets. These advanced computational approaches can be broken down into two categories, targeted and non-targeted, based upon the analytical objective (Figure 1.4). Targeted chemometrics refers to the use of decomposition algorithms to improve the identification and quantitation of pre-selected analytes of interest. Generally, if the analytes of interest are well-resolved in the chromatogram, or effectively so due to the selectivity provided by the detector employed, then advanced chemometrics is unnecessary. However, if the analyte is in a region with low R_s, traditional identification and quantitation efforts will be severely hindered. Chemometric decomposition of this region can be beneficial in extracting the pure signal for the specified analyte. On the other hand, non-targeted chemometric methods aim to categorize samples and discover compounds responsible for sample differentiation. Non-targeted chemometric techniques can be further categorized as unsupervised or supervised, depending upon if the method leverages class-based information based upon the experimental design. While this section will discuss these methods in their respective groups (Figure 1.4), these methods can be applied in any order depending upon the analytical objective. For example, the resulting quantitative information gained from targeted methods can be used to

build non-targeted models, or in contrast, targeted methods can be developed to identify and quantify analytes based upon the findings of non-targeted methods.

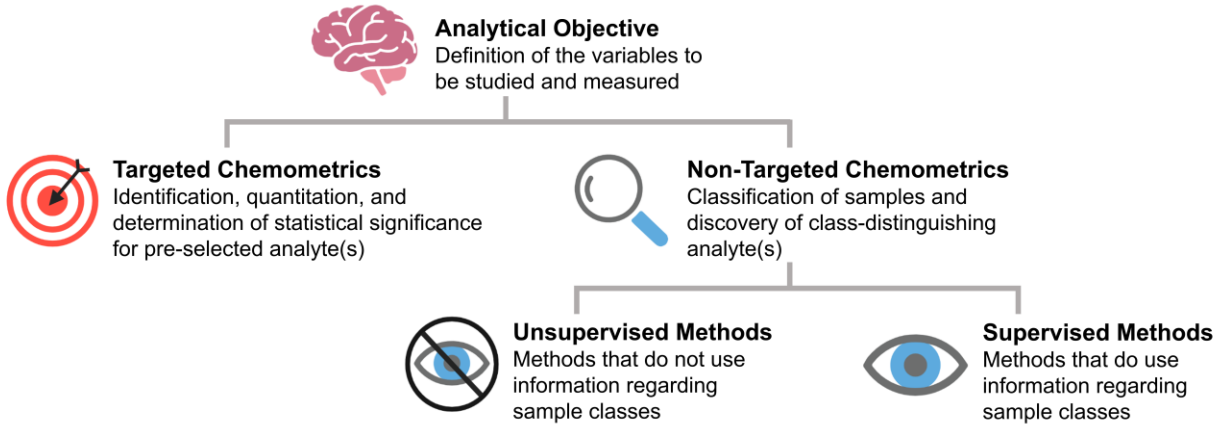


Figure 1.4. Overview of chemometric methods based on their approach and analytical goal.

1.3.1. Initial data analysis considerations

1.3.1.1. Chromatographic data structure

For any of these broad chemometric categories, application of specific methods will depend on the dimensionality of the data. Since this dissertation will focus on 1D-GC-MS and GC×GC-MS data analysis, the data structure for these chromatograms are illustrated in Figure 1.5. A single 1D-GC-MS chromatogram has a second-order data structure, which results in a 2D matrix with dimension representing t_{sep} and *m/z*. Meanwhile, the data structure for a GC×GC-MS is third order and can be described as a 3D cube, where the three axes describe the t_{sep} on both dimensions ($^1t_{sep}$ and $^2t_{sep}$) and *m/z*. Higher-order 1D-GC-MS and GC×GC-MS data can be achieved by analyzing multiple samples together. The order and quality of the chromatographic data produced is the most decisive step in selecting the appropriate chemometric method(s) for later analysis. However, while an instrumental design can produce higher-order data, its operation may not allow for the analyst to fully utilize second-order and third-order chemometric advantages. To realize these advantages, the data must be either bilinear or trilinear, meaning

that each data dimension (two for second-order data or three for third-order data) is linearly independent and analytes have sufficiently reproducible peak shapes, retention times, and concentration-dependent signals. Therefore, detector (and modulator for GC×GC) selection along with its impact on the data preprocessing workflow will ultimately influence the success of the chemometric analysis.

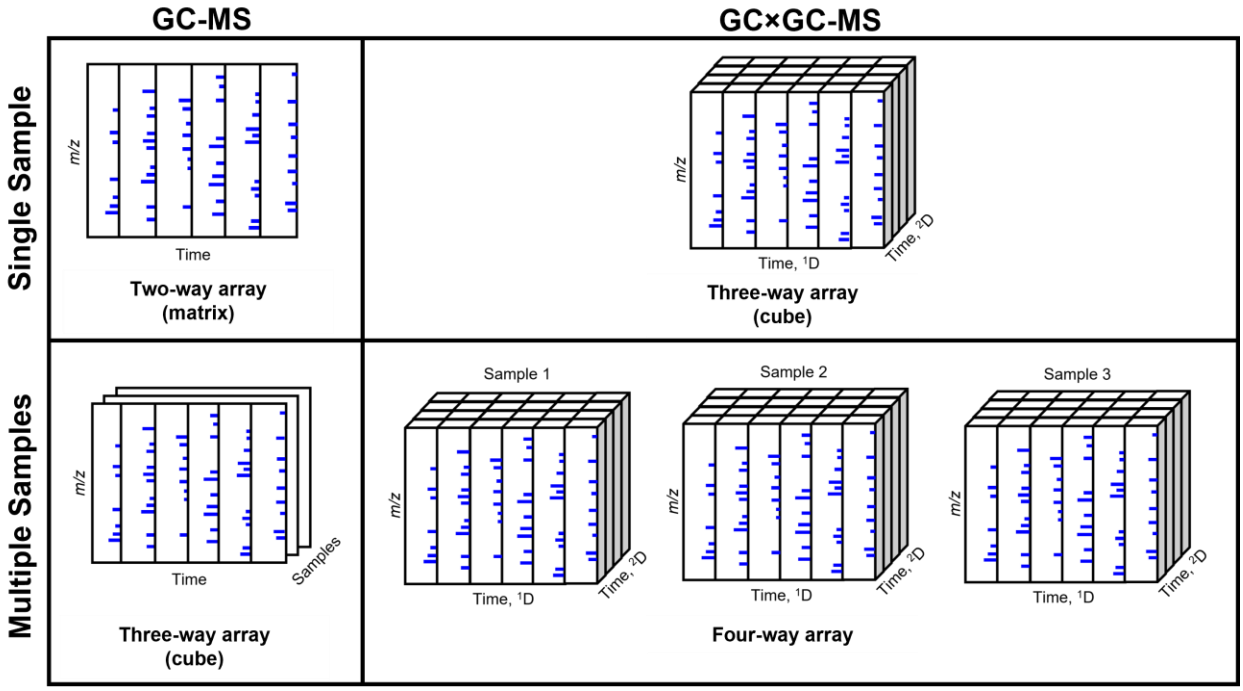


Figure 1.5. Schematic of the different dimensionalities for 1D-GC-MS and GC×GC-MS data. A single 1D-GC-MS and GC×GC-MS chromatogram has a second-order and third-order data structure, respectively. The dimensionality of the chromatographic data can also be increased by analyzing multiple samples simultaneously.

1.3.1.2. Chromatographic data preprocessing

Prior to analyzing 1D-GC-MS or GC×GC-MS data, some degree of data preprocessing is generally required to remove chemically irrelevant variations in the signal to improve chemometric performance. Baseline correction, smoothing, normalization, and retention time alignment methods are commonly used for data preprocessing. Low frequency detector noise (i.e., baseline drift) can be removed using baseline correction methods, which commonly subtract

a fitted curve from the entire chromatogram or sections of the chromatogram. Smoothing methods, such as a Savitzky-Golay filter, are used to reduce high frequency noise and increase the *S/N*. These two preprocessing methods must be carefully applied to prevent the loss of chromatographic signal and introduction of new artifacts, which can negatively impact chemometric performance. When comparing multiple replicates and/or samples, normalization and retention time alignment methods must be applied to reduce the inevitable variation from sample preparation and instrument operation. The use of internal standards or total area normalization, where the sum of the baseline corrected signal acts as the normalization factor, are the most common normalization methods. Retention time alignment programs ensure that variables that correlate to the same peak are correctly compared, and the bilinear (or trilinear) nature of the data is preserved. A variety of retention time alignment programs have been developed for chromatographic separations like piecewise alignment [58], correlation optimized warping [59], and dynamic time warping [60]. For data collected with a multivariate detector, algorithms can also use the collected spectra to improve retention time alignment results [61,62]. Another method to resolve misaligned chromatographic data is to average (i.e., bin) the data along the separation axis (or axes for GC×GC). Here, the appropriate bin size should be large enough to encompass the peak widths on both dimensions as well as the observed shifting [63]. As a result, proper binning increases the *S/N* while reducing the overall size of the data, which improves computational speed and performance. However, if the chromatogram is not appropriately binned, then a loss in chromatographic resolution can be observed.

1.3.1.3. Data analysis strategies for non-targeted chemometrics

Along with selecting the appropriate preprocessing methods for the data set, the analyst must also consider how the chromatographic data will be analyzed with chemometrics. Herein,

we broadly define three approaches for chemometric data analysis of chromatographic data: pixel-based, peak table-based, and tile-based. The selection of any of these data analysis platforms is dependent on both the size of and preprocessing methods selected for the chromatographic data set. The presence of these instrumental artifacts, also referred to as false positives, during analysis burdens the workflow and can hinder the identification of true chemical differences. Common instrumental artifacts that can be falsely discovered by non-targeted chemometric methods are random fluctuations in the baseline noise and retention time misalignment between samples. Therefore, the appropriate data analysis approach should reduce the chemometric discovery of instrumental artifacts.

Pixel-based analysis directly applies chemometrics on every single intensity measurement (i.e., pixel) in each chromatogram [64]. This approach ensures that the original structure and information in the chromatogram is preserved throughout the analysis. The results from pixel-based data analysis can then be visualized using the structure of a chromatogram, supporting chemical interpretation. However, since chemometrics is performed using the raw data, preprocessing methods such as baseline correction and retention time alignment must be carefully chosen to minimize the discovery of false positives. Misalignment can be a key obstacle to pixel-based analysis since the intensity measured at every data point would not correspond to the same feature throughout the data set. Along with preprocessing concerns, the size of the chromatogram (or data set) must also be carefully considered since pixel-based analysis can be computationally expensive. For example, the file size for a single GC×GC-MS chromatogram can be anywhere from a couple hundred megabytes (MBs) to hundreds of gigabytes (GBs) depending on the separation time, choice of MS, and specifications of the MS (i.e., collection frequency, *m/z* range, etc.). Hence, the pixel-based analysis of multiple GC×GC-

MS chromatograms only increases the computational burden. Therefore, application of pixel-based chemometrics on specific regions of the chromatogram or pre-selected *m/z*s is necessary to improve computational performance and speed.

Compared to pixel-based analysis, both peak table-based and tile-based approaches inherently provide a degree of data reduction prior to chemometric analysis. Peak table-based approaches rely on identifying and quantifying "every" peak above a given *S/N* across all chromatograms [53]. Peak table generation for a chromatographic data set can easily be performed within the instrumental software suite or in commercial packages [53,65]. Generally, peak tables for each chromatogram are produced first, and then the entries between the peak tables are aligned to one another using predetermined match criteria. For GC×GC-MS data, Bean et al. advised that entries should be aligned if their ^{1}D and ^{2}D retention times are within ± 1 P_M and 100 ms, respectively, and if the match value (MV) comparing their mass spectra is greater than 600 [53]. Chemometric data analysis is then performed on the final peak table, which contains the integrated peak signal for all peaks resolved in the data set. Generating high quality peak tables for chemometric analysis is highly dependent on the quality of the chromatographic data. Retention time shifting and overloaded peaks in the data set should be minimized to ensure that the signal for a peak is not split among multiple entries in the final table. However, even when processing high quality chromatograms, values for peaks with low signal can be missing from the final table due to the *S/N* filter utilized by the peak finder in the software. These missing values can lead to erroneous conclusions about the data set and the analytical objective at-large [66]. Strategies to overcome these missing entries involve either removing peaks from the table whose signal is missing in more than 20 % of the samples and/or using an imputation approach [66].

Lastly, a tile-based approach was developed by Synovec and co-workers to address common drawbacks associated with pixel- and peak table-based analyses, such as mitigate run-to-run retention time misalignment, reduce computational load, and improve the discoverability of analytes with a low *S/N* [67,68]. Figure 1.6 demonstrates the application of tiling on GC×GC-MS chromatograms. Using preprocessed data, this approach divides the GC×GC chromatograms into small, rectangular sections (i.e., tiles) and sums together (i.e., bins) the signal captured within those sections. Generally, the tile size should be large enough to encompass the average peak width and any additional retention time shifting [63,67,68]. For situations involving a large degree of misalignment and/or within-class variability, a tile size larger than the general recommendation may be beneficial in reducing the discovery of false positives [69]. This tiling process is then repeated using four grid schemes, which are offset by half the original tile size in either or both dimensions, to ensure that every peak in the chromatogram is best encapsulated by one of the four grid schemes (Figure 1.6). Chemometric analyses are then performed on a per-*m/z* basis using the tiled GC×GC chromatograms. Note, this four-grid tile scheme can also produce multiple results for a single peak, which are called redundant hits. Therefore, this redundancy is removed using a "pinning and clustering" algorithm. This algorithm first locates the 2D maximum for each tile (i.e., the pin location) and then consolidates all pins with similar retention times down to a single pin (Figure 1.6). The finalized pin locations can then be used to further investigate the chromatographic data using the original high-fidelity data at the pixel-level. While this tiling methodology is discussed in the context of GC×GC-TOFMS data, it is important to note that this method has recently been adapted for comparative analyses involving 1D chromatography coupled to MS detection [70] and comprehensive three-dimensional gas chromatography (GC^3) with TOFMS detection [71].

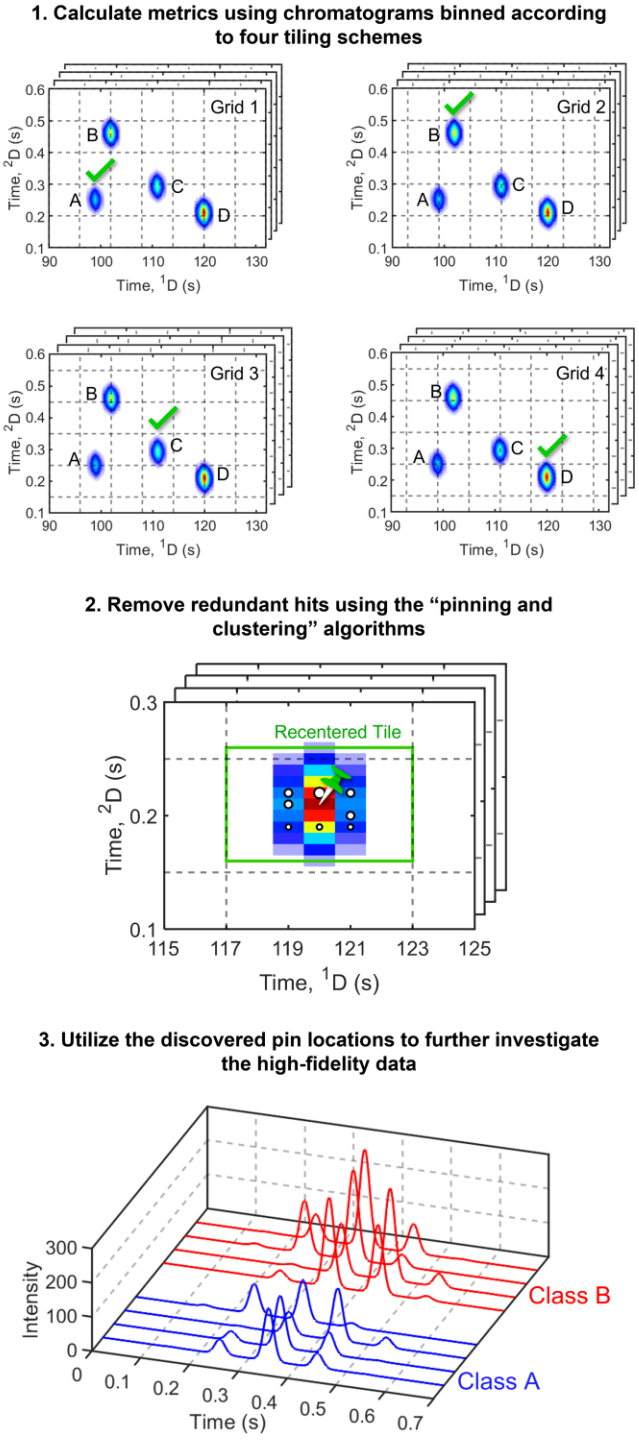


Figure 1.6. Illustration of tile-based chemometric analysis for GC×GC-MS data. (1) First, the data are binned into four grid schemes to optimally capture each analyte. (2) Then, redundant hits from the multiple grids and/or peak splitting are removed by "pinning and clustering" and the tile is adjusted around the peak. (3) Finally, the analyst can unfold the tiles back to the original high-fidelity data.

1.3.2. Targeted chemometric methods

The primary goal for any separation is to identify and quantify analytes responsible for the similarities and differences in a data set. In targeted studies, the identity of these analytes of interest is known beforehand, and the separation is designed to chromatographically resolve each targeted compound to the greatest extent possible. After the data is collected, the identity of the target analytes is confirmed via spectrum library matching and/or retention time indexing with analyte standards. Then, analyte concentration is determined using the standard addition method, external standards, or internal standards and the measured peak heights/areas [72,73]. However, the experimental design cannot always be optimized for all the target compounds of interest to be fully resolved, causing overlapped interferent signals to challenge identification and quantitation. Therefore, chemometric decomposition methods can be used to obtain pure chromatographic peak profiles and spectra. It is important to note that chemometric decomposition has also been referred to as deconvolution in the literature. This section will focus on two popular decomposition methods: multivariate curve resolution-alternating least squares (MCR-ALS) and parallel factor analysis (PARAFAC). The operation of both methods is similar even though they have different data structure requirements. For example, both methods are traditionally applied to relatively small regions of the chromatogram instead of the entire chromatogram to lighten their computational load. Both MCR-ALS and PARAFAC also require the analyst to provide an estimate of the number of mixture components separated in the selected time window (i.e., the rank of the data). Generally, the number of components is taken as the number of analytes present plus additional component(s) for the baseline/background noise. This estimate of the rank will remove background and noise from the pure component profiles without the need of baseline correction steps. These methods then leverage information in each data dimension to

mathematically resolve target and interferent signals. To develop an accurate decomposition model, the experimental design must ensure the chromatograms adhere to either a bilinear or trilinear data structure.

1.3.2.1. Multivariate curve resolution-alternating least squares (MCR-ALS)

MCR-ALS is a bilinear decomposition method, which extracts the pure component information for each dimension of second-order data [74–76]. Given the bilinearity requirement, MCR-ALS can be applied to both 1D-GC-MS data and GC×GC-MS data. To ensure bilinearity with 1D-GC-MS chromatograms, the data must be aligned to minimize retention time shifting. Meanwhile, prior to applying MCR-ALS to GC×GC-MS, the dimensionality of the data must be reduced prior to MCR-ALS. This data reduction can be achieved by analyzing individual modulations (i.e., the ^{2}D separation) or unfolding the time dimension (i.e., concatenating each ^{2}D separation together) while maintaining the spectra dimension. A benefit of applying MCR-ALS to chromatograms collected with multivariate detection is that alignment is not necessary because data bilinearity is supported by the reproducibility of the spectra dimension [77]. The MCR-ALS model can also be extended to simultaneously analyze multiple samples or replicates. For these higher-ordered arrays, the time dimension for each sample would be unfolded and then those samples would be augmented together along the time axis (Figure 1.7).

The MCR-ALS model is represented as

$$\mathbf{X} = \mathbf{R}\mathbf{S}^{\mathbf{T}} + \mathbf{E} \tag{1.8}$$

where **X** is the chromatographic data matrix, **R** and **S** are matrices containing the pure instrumental responses, and **E** is a matrix containing the residual errors [74–76]. For 1D-GC-MS and GC×GC-MS data, the matrix **R** generally represents the resolved chromatographic elution profiles (i.e., time dimension) for each modeled component. Likewise, the matrix **S** normally

contains the pure *m/z* for each component in the model. Since MCR-ALS is an iterative method, the algorithm will alternate between the results in **R** and **S** to minimize the errors in **E**. Figure 1.7 illustrates the format of a two-component MCR-ALS model to analyze either a (A) single chromatogram or (B) multiple chromatograms simultaneously. The ease of obtaining pure information from the experimental data is dependent on the number of estimated components in the subsection of the chromatogram along with the *S/N* of the target analyte, its relative intensity, and extent of overlap with interferents. Using the model outputs, the pure elution profile and spectrum for each component can be used for quantitation and identification, respectively.

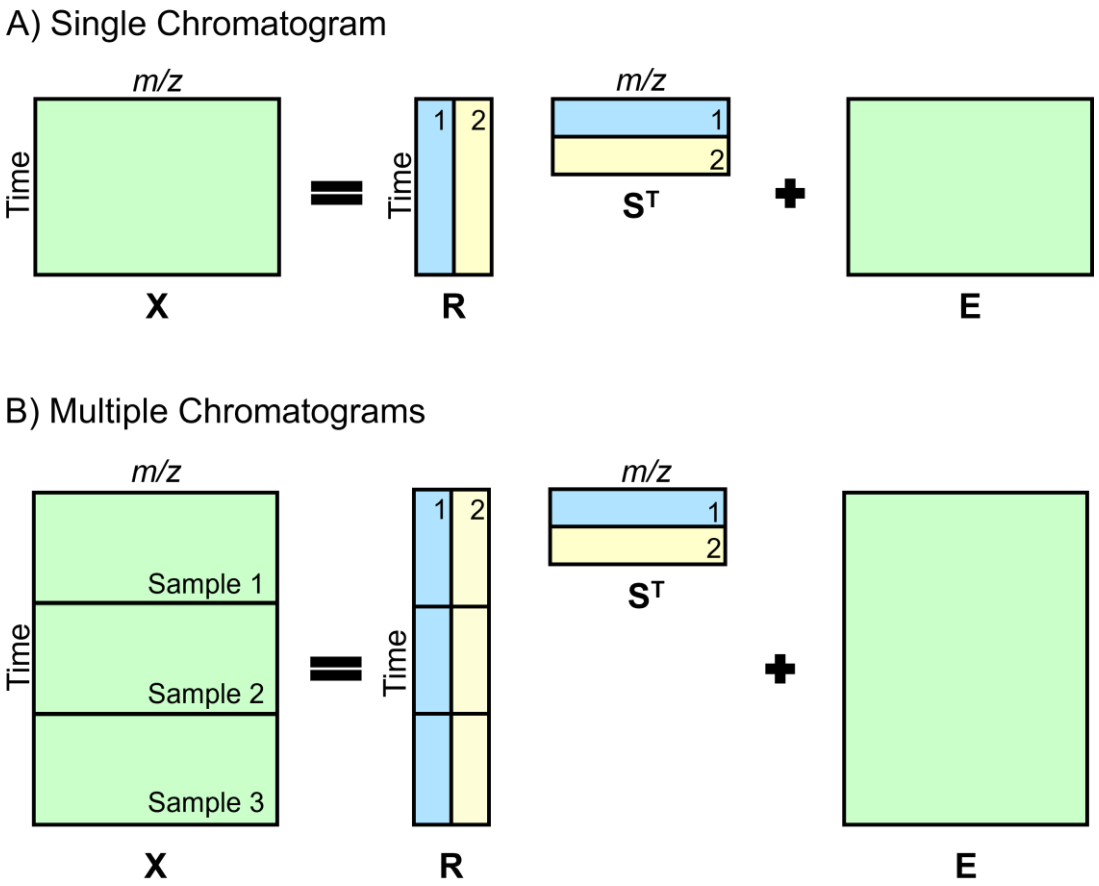


Figure 1.7. Illustration of a two-component MCR-ALS model for either 1D-GC-MS or GC×GC-MS data. The model decomposes the chromatographic data (**X**) into pure chromatographic (**R**) and mass spectral (**S**) profiles for each component. Components 1 and 2 are highlighted in blue and yellow, respectively. MCR-ALS models can be constructed for (A) single chromatograms or (B) chromatographic data sets with multiple samples.

While MCR-ALS is a flexible decomposition model for chromatographic data, it is possible that different solutions can be produced for the same matrix input and those solutions can fit the data equally well. This uncertainty is referred to as "rotational ambiguity" and the extent of this uncertainty can be evaluated by finding all possible, feasible solutions [78,79]. Proper initialization and selection of constraints can reduce the number of possible solutions, improving the fit of the MCR-ALS model. MCR-ALS initialization refers to providing the model of an initial estimate of a data dimension for each component. Typically, for chromatographic applications, initial estimates are provided for the spectrum of each modeled component. These initial estimates can either come from prior knowledge (e.g., the pure spectrum for the target analyte) or algorithms designed to select the most dissimilar spectra in the original data. The most common initialization methods include simple-to-use self-modeling analysis (SIMPLISMA) [80], orthogonal projection approach (OPA) [81], and key set factor analysis (KSFA) [82]. Both OPA and KSFA can be performed in an iterative manner for further refinement of the spectra to be used as initial estimates [83–85]. Additionally, constraints place mathematical conditions on the fit of **R** and **S** during the iterative optimization of the MCR-ALS model. The most common constraints for chromatographic data are non-negativity, ensuring the elution profiles have non-negative concentrations, and unimodality, ensuring only one peak maximum per component. Defining regions in the chromatographic data with an absence of analytes (i.e., local rank constraints), concatenating replicates prior to decomposition, or using hard modeling can also be implemented to mitigate rotational ambiguities [79]. Additionally, application of a trilinearity constraint can be used to obtain essentially the same unique solution as higher-order decomposition models like PARAFAC [79].

1.3.2.2. Parallel factor analysis (PARAFAC)

PARAFAC is a trilinear decomposition method, which can extract the pure instrumentally obtained responses from third- or higher-ordered data sets [86]. PARAFAC is commonly performed on GC×GC-MS chromatograms since the data structure is naturally third-order (^{1}D time × ^{2}D time × *m/z*). Compared to MCR-ALS, PARAFAC is advantageous for decomposition of GC×GC-MS separations since it does not require a reduction in data dimensionality and the final solution in the model is unique. However, data submitted for PARAFAC modeling must be sufficiently trilinear whereas MCR-ALS only requires bilinear data. The strict trilinear structure condition requires all the ^{2}D peaks for a single ^{1}D peak to be reproducible in terms of peak shape, width, and retention time (or at least not deviate greatly) and chemically selective information must be present in at least two of the data dimensions [87].

Given a chromatographic data cube **X**, consisting of elements x_{ijk} and F number of components (i.e., the rank of the data), the PARAFAC model can be expressed as

$$x_{ijk} = \sum_{f=1}^{F} a_{if} b_{jf} c_{kf} + e_{ijk} \quad (1.9)$$

where a_{if}, b_{jf}, and c_{kf} are the elements of matrices **A**, **B**, and **C** containing the pure instrumental responses for each component and e_{ijk} are the elements of a three-way array, **E**, representing the residual error [86]. For GC×GC-MS, the columns of the matrices **A**, **B** and **C** (the loadings) correspond to the chromatographic profiles in both dimensions (^{1}D and ^{2}D) and the *m/z* for each component modeled, respectively. Figure 1.8 demonstrates the construction of a two-component PARAFAC model to analyze a single GC×GC-MS chromatogram. The model is achieved by using initial estimates for two dimensions and then applying alternating least squares to fit the remaining mode to obtain the solution where the residuals, e_{ijk}, are minimized.

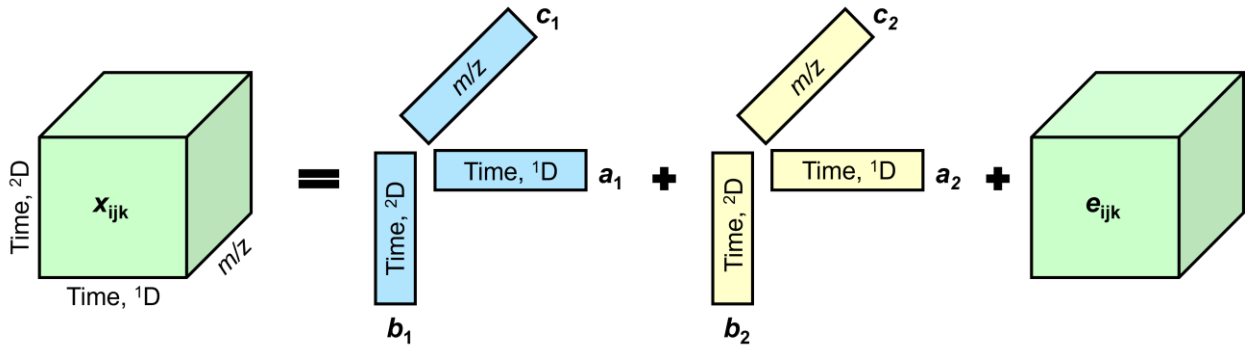


Figure 1.8. Illustration of a two-component PARAFAC model for GC×GC-MS data. The model decomposes the chromatographic data (**X**) into pure chromatographic (**A and B**) and mass spectral (**C**) profiles for each component. Components 1 and 2 are highlighted in blue and yellow, respectively.

Selection of the number of components to model, initialization and stoppage values, and constraints are all necessary to execute PARAFAC modeling. Like MCR-ALS, the number of components to model should equal the number of suspected analytes plus one or more for the background contributions. If too many components are used in the PARAFAC model, then the computational speed decreases and true analyte signals can be modeled by multiple components (i.e., splitting). Hoggard and Synovec defined the appropriate number of components to model as one fewer than the PARAFAC model with the observed splitting [88]. This method successfully created PARAFAC models for target analytes across a wide range of signal intensities (overloaded to low *S/N*) and could be applied in an automated fashion [88]. Split-half experiments can also help determine the correct number of components to model [89]. Here, the chromatographic region is divided into two sections and PARAFAC models are created for both sections. If the number of components is selected correctly, the same loadings should be evident in the models of both data sets. Typically, random values or initial estimates from trilinear decomposition are used for initialization while the minimum of the residuals array (**E**) is a stoppage criterion. Unimodality and orthogonality constraints can help stabilize the solution while non-negativity ensures the loading vector should have positive signal [86].

Along with identification and quantitation, PARAFAC can also be used to evaluate the trilinearity (or higher) of the chromatographic data structure [90–92]. The magnitude of retention time shifting between modulations to ^{2}D peak width, referred to as the trilinearity deviation ratio (TDR), can predict the accuracy of PARAFAC models for quantitation [87,93]. Application of PARAFAC to non-trilinear data was shown to cause a negative bias, where true analytical signal that does not fit the model has been removed [87,93]. Furthermore, the trilinear nature can be assessed by comparing the loadings from PARAFAC to the experimental data by calculating two metrics, the lack-of-fit (*LOF*) and percent of explained variance (R^2) [90–92]. Ideally, if the chromatographic data is trilinear, then the measured *LOF* and R^2 will be 0 % and 100 %, respectively. Note, experimental conditions have the greatest influence on the trilinear (or bilinear) nature of the chromatographic data. For example, Prebihalo et al. demonstrated that GC×GC-TOFMS data is sufficiently trilinear (i.e., small TDRs and PARAFAC quantitation errors) with a small P_M (~1-2 s) compared to relatively longer P_M (~5-8 s), which are typically used [93]. In cases where the experimental design was not optimized to ensure a trilinear data structure, retention time alignment algorithms should be used to make the data more amenable to PARAFAC [62,90,94]. For instance, Allen and Rutan demonstrated that alignment improved the accuracy and reproducibility of quantitative PARAFAC models for phenytoin in wastewater samples [62]. PARAFAC2 can also be used to analyze three-way chromatographic data (e.g., 1D-GC-MS data sets) that do not follow a trilinear nature. As a modified version of PARAFAC, PARAFAC2 is less sensitive to misalignment, the main deviation from trilinearity, while still producing unique solutions for all three data dimensions [95].

1.3.3. Unsupervised, non-targeted chemometric methods

While targeted chemometric methods are beneficial in the identification and quantitation of previously known and anticipated analytes of interest, non-targeted approaches seek to discover relevant chemical features that describe the similarities and/or differences across multiple chromatograms. Non-targeted chemometric methods can be described as either supervised or unsupervised, where supervised methods depend upon *a priori* knowledge of sample classification. Supervised approaches (discussed later in *1.3.4. Supervised, non-targeted chemometric methods*) are appropriate for handling classification and regression problems since these methods leverage class labels. In contrast, unsupervised models do not require knowledge of class memberships. Therefore, unsupervised approaches are suitable for exploratory data analysis, where the user aims to discover patterns and detect outliers in the data set. These unsupervised, non-targeted methods are typically the first step in a chemometric workflow because they are simple, computationally inexpensive, and provide visualization of the main attributes of the data. This section will cover two common unsupervised techniques for chromatographic data analysis: principal components analysis (PCA) and *k*-means clustering. PCA provides identification of features that accurately represent relationships between samples, whereas *k*-means clustering discovers the inherent groupings hidden in unlabeled data.

1.3.3.1. Principal components analysis (PCA)

PCA is quite possibly the most applied exploratory data analysis technique because it reduces the chromatographic data down to only the variables that represent the variation and correlations in the data set. This data reduction is achieved by projecting the possibly correlated variables (i.e., peaks) in the data onto a new set of linearly uncorrelated variables called principal components (PCs) [96]. After the orthogonal transformation, these PCs are then ranked in

descending order of explained variance. Therefore, PC 1 explains the maximum variability in the data, PC 2 explains the maximum variance not explained by PC 1, and so on. This process continues until all the variance in the data set has been explained or until an algorithmic stopping point has been reached [97]. In practice, only the first couple PCs that explain a proportion of the total variance in the model will be kept and utilized for interpretation.

The output of PCA is a decomposition model, which can be described as

$$\mathbf{X} = \mathbf{TP} + \mathbf{E} \tag{1.10}$$

where **X** is the original chromatographic data matrix, **T** is the scores matrix, **P** is the loadings matrix, and **E** represents the unaccounted signal that remains. A visual illustration of PCA model generation on unfolded GC×GC-MS chromatograms is provided in Figure 1.9. The data matrix (**X**) must be a two-way array, where the rows represent the sample dimension and columns are the variables. The variables can either be the data points for the completely unfolded chromatograms or tabulated peak areas. Note, multiway PCA [98] develops the same decomposition model as PCA but for third-order data, preserving chemical information that can be lost when reducing the dimensionality of 2D chromatographic data. Examination of the scores and loadings matrices can provide information on the similarities and/or differences between samples and chemically relevant analytes, respectively. The scores matrix (**T**) describes the coordinates for each sample on the PC axis and the loadings matrix (**P**) highlights the peaks responsible for the variation described by each PC. Plotting the scores on PC 2 versus the scores on PC 1, termed a scores plot, illustrates the relationship between samples in a data set. Ideally, similar samples should have similar scores while dissimilar samples should be separated from one another on the scores plot. The separation between these clusters can be quantified using a variety of metrics such as a degree-of-class separation (DCS) metric [58,63], the Mahalanobis

distance [99], and construction of confidence ellipses [100,101]. Furthermore, investigation of the loadings for each PC can determine the peaks that are responsible for the sample separation on the scores plot. For a given PC, peaks with positive loadings are more abundant in samples with positive scores while peaks with negative loadings are more abundant in samples with negative scores.

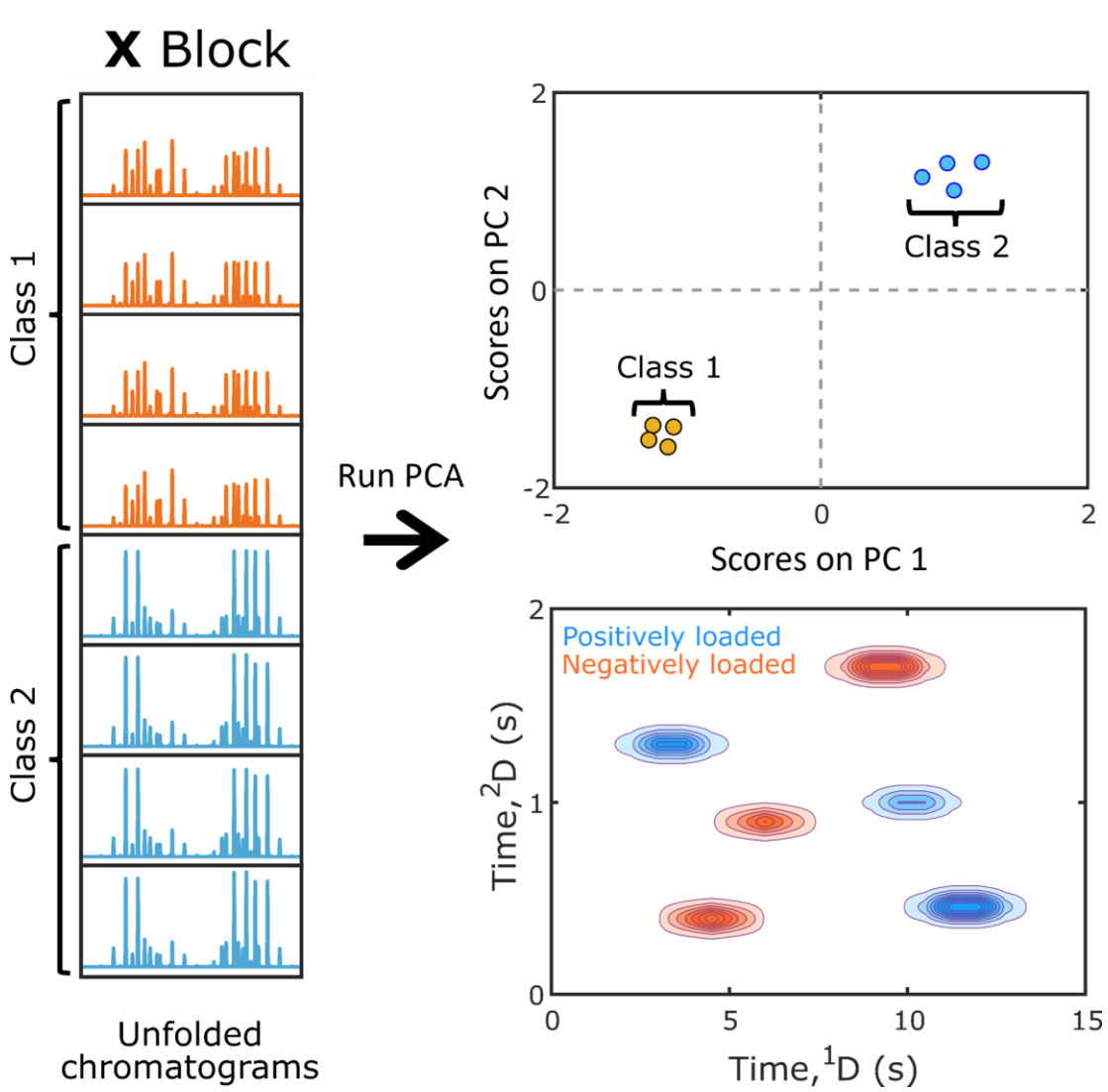


Figure 1.9. Schematic illustrating PCA for GC×GC-MS data with two classes known *a priori*. The **X**-block contains the chromatograms in their vectorized form. Note, the class labels for the data do not need to be known prior to PCA. Following PCA, the model outputs both scores and loadings for each PC. The scores plot shows the coordinates for each sample on each PC. The loadings highlight the features that are both positively (blue) and negatively correlated (red) to the given property.

Since PCA is extremely sensitive to all sources of variance, applying preprocessing methods is essential to attaining chemically meaningful results. Along with using baseline correction and normalization techniques, retention time shifting should also be reduced. If chromatographic misalignment is not corrected for, then the first few PCs can capture the variance due to shifting instead of the sample-related variances that are of interest. The loadings plot for each PC will also show first derivative Gaussian-like signals instead of chromatographic peaks due to retention time shifting [102]. Therefore, a strategy to mitigate chromatographic misalignment must be employed prior to PCA. The use of peak tables is a common approach to overcome retention time shifting. These tables are typically generated by commercial software, which attempts to identify and quantify the analytes present in each chromatogram before aligning the tables. However, this approach may not be successful for highly saturated chromatograms where multiple chemical species overlap. Therefore, the analyst may want to either apply an alignment algorithm to the data set or bin the data to overcome misalignment. Application of a retention time alignment algorithm to pixel-level data (i.e., every data point in the unfolded chromatograms) was shown to increase the DCS between different gasoline samples [58]. Binning the pixel-level data can also minimize chromatographic misalignment while increasing the *S/N*. Sudol *et al*. showed that a maximum DCS between two fuel classes on a scores plot occurs at an optimal level of binning [63]. However, when examining the five adjacent fuel pairs in the study, each one had a different optimum bin size due to the degree of chemical differences between the fuels [63]. This result indicates that the analyst must select a bin size that balances the *S/N* improvement while minimizing misalignment and maintaining chemical selectivity.

1.3.3.2. Partitional clustering analysis

A partitional clustering algorithm can also be used to visualize and quantify the similarities and differences between chromatographic samples. These algorithms work by simultaneously assigning all samples into *k* clusters, where *k* represents the number of specified groups in the data set. Then, these algorithms iteratively relocate the samples between clusters until the sum-of-squared distances is minimized. The most popular partitional clustering algorithm is *k*-means clustering, which defines each cluster centroid as the average of all samples assigned to that cluster [103]. The algorithm randomly assigns *k* cluster centroids and the distances between sample and cluster centroids are calculated. Either the Euclidean or Manhattan distance can be used for this calculation. During the cluster assignment and centroid update steps, the samples are grouped to their closest cluster centroid and the algorithm determines the new cluster centroids. The algorithm then recalculates the sample-to-centroid distances and repeats the cluster assignment and centroid update steps. This method is repeated until cluster memberships do not change, or a maximum number of iterations is reached. Due to the random selection of centroids, the resulting cluster assignments are not reproducible between algorithm runs. Therefore, the *k*-means algorithm is commonly performed multiple times with different initial centroids and the model with the smallest sum-of-squared distances is selected as the appropriate model [104]. Other variations of *k*-means clustering, which heuristically select the initial centroids or limit the cluster centroids to be a member of the cluster, have also been proposed to improve model reproducibility [104,105].

Appropriate selection of the number of clusters, *k*, to model is imperative to achieving useful cluster assignments. In practice, cluster assignments at different values of *k* are compared using a clustering validity index [104,106]. Numerous index calculations have been described in

the literature with the goal of quantifying within-cluster compactness and between-cluster separation [106]. However, the silhouette index [107] and Davies-Bouldin index [108] are typically used. The silhouette index provides a measure of how similar a sample is to others in its own cluster relative to other samples in a neighboring cluster. The resulting metric is termed a silhouette value, which ranges from -1 (not well clustered) to 1 (well clustered). The appropriate number of clusters, *k*, can be determined by selecting the clustering solution that had an average silhouette value closest to 1. Similarly, the Davies-Bouldin index measures the ratio of the within-cluster variance to between-cluster distances. The optimal clustering solution has the smallest Davies-Bouldin index value.

1.3.4. Supervised, non-targeted chemometric methods

While unsupervised approaches are appropriate for initial investigations into a chromatographic data set, supervised approaches are well suited for studying cause and effect experiments by leveraging *a priori* information. Supervised algorithms utilize target variables like class labels or independently measured sample properties to discover features, build regression models, and/or classify samples. Feature discovery, also known as feature selection, finds a subset of the original chromatographic data that is highly correlated with the target variable(s). For chromatographic data sets, Fisher ratio (F-ratio) analysis is typically used to discover class-distinguishing analytes. It is important to note that unsupervised methods like PCA are commonly used to visualize the results obtained from non-targeted, supervised feature selection methods. Along with identifying significant analytes (feature selection), methods used for property prediction and sample classification fall under the umbrella of supervised analysis techniques. Note, in this context property refers to either a chemical or physical quantity that was collected separately from the chromatographic data set. The most common property prediction

method is partial least squares (PLS) regression, which develops a multivariate calibration model to discover which analytes correlate to the sample property that is being modeled. The details of both F-ratio analysis and PLS regression are discussed in the following sections.

1.3.4.1. Fisher ratio (F-ratio) analysis

F-ratio analysis is a popular feature selection technique for chromatographic data because it inherently provides a degree of data reduction, focusing the overall data analysis before performing further targeted and non-targeted chemometric methods. This feature selection method utilizes the analysis of variance (ANOVA) statistical hypothesis test, which compares the variance of observations and discovers significant differences between groups. The total variance, defined as the squared standard deviation, can be partitioned into two contributions: variance *between* classes of samples and *within* classes of samples. The between class (BC) variance, which describes how each class mean varies from the grand mean, is defined as

$$\sigma_{\mathrm{BC}}^2 = \frac{1}{k-1}\sum(\bar{x}_{\mathrm{i}} - \bar{x})^2 n_{\mathrm{i}} \tag{1.11}$$

where k is the number of classes, n_i is the number of measurements in the ith class, $\overline{x}_i$ is the mean of the ith class, and $\overline{x}$ is the grand mean. The within class (WC) variance, which indicates how much each measurement varies from its class mean, is

$$\sigma_{\mathrm{WC}}^2 = \frac{1}{N-k}\sum\sum(x_{ij} - \bar{x}_i)^2 \tag{1.12}$$

where N is the total number of measurements, and x_{ij} is the jth measurement of the ith class. Finally, the F-ratio is then obtained by taking the ratio of these two quantities:

$$F - ratio = \frac{\sigma_{\mathrm{BC}}^2}{\sigma_{\mathrm{WC}}^2} \tag{1.13}$$

The results from F-ratio analysis are compiled in a "hit list," which ranks the F-ratio values in descending order. The analyst then mines the hit list in a top-down approach, identifying and quantifying peaks with larger F-ratios since a high F-ratio generally corresponds

to class-distinguishing analytes. A class-distinguishing analyte is an analyte whose concentration is statistically different between classes, which is typically based upon a *t*-test having a *p*-value < 0.05 (95% confidence limit). Thus, the results of the *t*-test show that the concentration ratio between classes sufficiently differs from one. These class-distinguishing analytes are commonly referred to as true positives. However, this data mining approach can be hindered by the presence of both false positive and negatives. A false positive refers to the discovery of an analyte that is not statistically different between classes while a false negative is the inability to discover a class-distinguishing analyte. The presence of false positives and/or negatives can be present in all implementations of F-ratio analysis (peak table, pixel-based, and tile-based). Therefore, effective preprocessing strategies must be applied prior to analysis to reduce the false discovery rate (discussed in *1.3.1.3. Data analysis strategies for non-targeted chemometrics*).

Even with proper mitigation of false positives, through appropriate preprocessing, discovery of significant class-distinguishing analytes by F-ratio analysis can be challenging due to any natural variation present in the data set and/or the number of hits "discovered". First, for multi-class metabolomics data with inherent within-class variance up to ~ 50 %, modifications to the F-ratio calculation shown in Eq. 1.13 have also been explored [109–111]. For instance, control-normalized F-ratio analysis uses only the within-class variance from the control group in the denominator of Eq. 1.13 [109,110]. For data sets with a large within-class variance in all sample classes, F-ratios can be calculated using the lowest within-class variance among all classes, termed minimum variance optimized (MVO) F-ratio analysis [111]. Second, the use of an F-ratio threshold can minimize analysis time during the data mining process. Initially, the F-critical value was utilized to trim the hit list [58,112]; however, it was determined that these cut-offs were too low and kept many false positives in the hit list [68,113,114]. Subsequently,

combinatorial null distribution analysis was utilized as a statistical means to determine an appropriate F-ratio threshold at a given null probability and confidence level [68,113,114]. Recently, Sudol et al. proposed using the *p*-value, either from a *t*-test or one-way ANOVA, as a means for determining the F-ratio threshold [39]. Ultimately, both methods improve the discovery, identification, and quantitation of class-distinguishing analytes in F-ratio analysis.

1.3.4.2. Partial least squares (PLS) regression

Partial least squares (PLS) analysis is a multivariate regression method, which correlates the information in a data matrix (**X**) to information in another matrix (**Y**), which may be a single column vector (i.e., PLS1), or a multi-column matrix (i.e., multiway PLS; n-PLS) [115]. PLS is often used to predict some property in **Y** that is difficult or expensive to obtain by the reference method, using chromatographic data (**X**). In essence, PLS analysis is based on performing PCA individually on **X** and **Y**. In PCA, the direction of the loadings in **X** and **Y** maximizes the variance within each of the respective matrices. However, in PLS, the direction of loadings of **X** and **Y** is chosen to maximize the covariance between these two matrices. Analogous to PCs, the variation in the **X**-block is described by a series of orthogonal linear latent variables (LVs). A visual illustration of PLS model generation and optimization on unfolded GC×GC chromatograms is provided in Figure 1.10. Note, PLS can be a computationally expensive technique and thus, is seldom performed on pixel-level chromatographic data. Therefore, data reduction techniques like removing uninformative chromatographic regions or *m/z* values, binning [116,117], and feature selection (e.g., F-ratio analysis) [118] can be used before constructing the **X**-block.

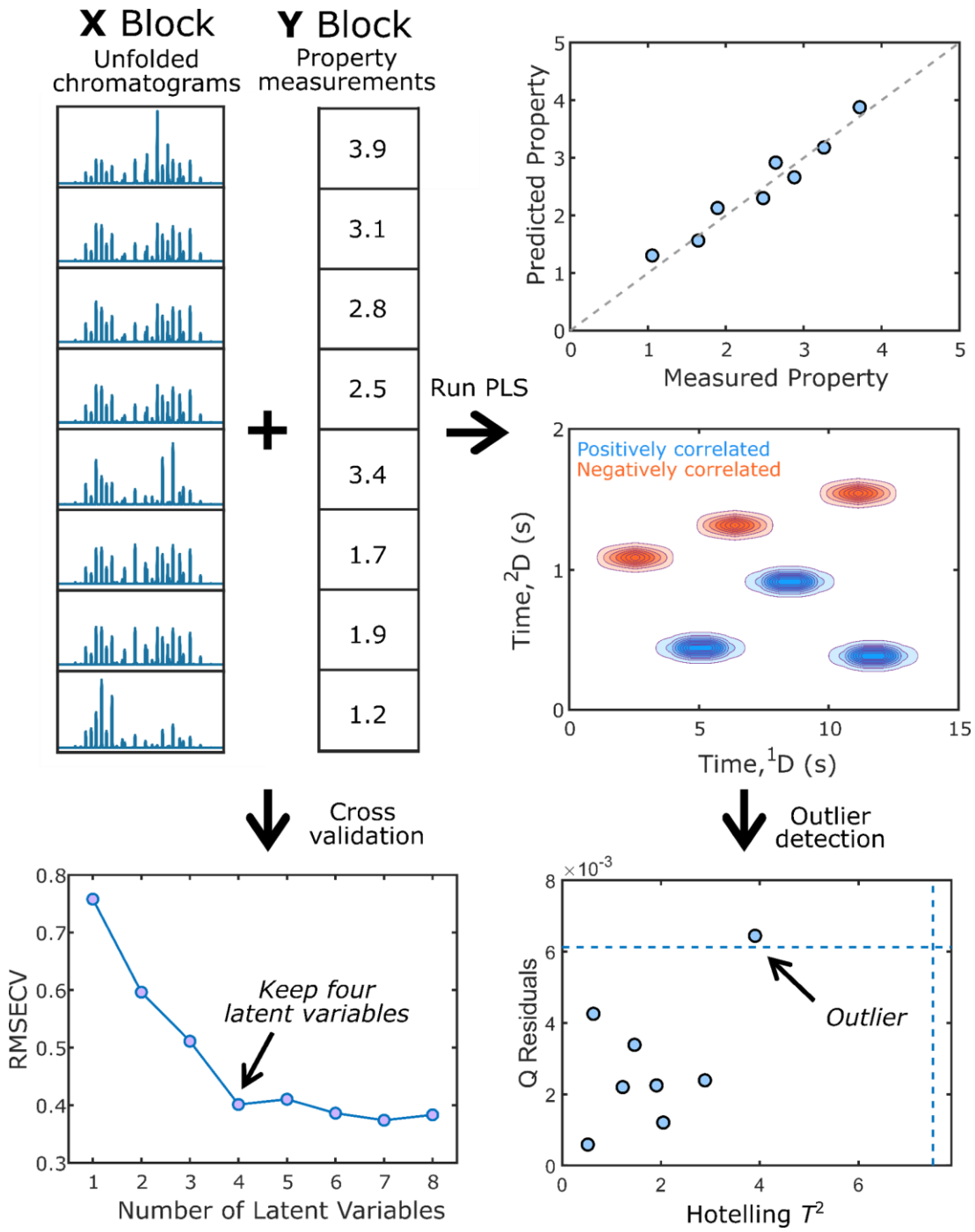


Figure 1.10. Schematic illustrating PLS regression analysis for GC×GC-MS data. The **X**-block contains the chromatograms in their vectorized form, and the **Y**-block contains the corresponding property measurements. Following PLS, a regression model and loadings (or linear regression vectors; LVRs) are generated. The regression plot shows the predicted property values from the model versus measured property values. The LVRs highlight the features that are both positively (blue) and negatively correlated (red) to the given property. Cross-validation is performed to determine the number of latent variables retained in the PLS model. Furthermore, a plot of the *Q* residuals versus Hotelling T^2 can be used for outlier detection.

The number of LVs to include is an important step in the development of a PLS model. This is generally determined via leave-one-out-cross-validation (LOOCV), wherein one row of **X** (e.g., an unfolded chromatogram) is excluded from the model, and the model is rebuilt. After repeating this for every row of **X**, the root-mean-square error of cross-validation (RMESCV) is computed, which measures the difference in the cross-validation predicted value of the samples and their measured values from **Y**. Then, the number of LVs to include is selected from the model with the lowest RMESCV, or after the change in RMESCV upon adding additional LVs becomes negligible (Figure 1.10). However, LOOCV is computationally expensive for large data sets and can result in model overfitting. Hence, the cross-validation method may employ one or more sub-validation experiments, where a subset of samples, rather than a single sample, is removed from the **X**-block to generate a validation set. Examples of sub-validation methods include Venetian blinds, contiguous blocks, and random subsets [119]. These sub-validation methods differ in how the samples to remove from **X** are selected. The proper selection of a sub-validation model(s) depends on the nature of the data set and the analysis goals.

The primary outcome of a PLS model is a regression plot showing the correlation of the predicted property to the measured property (Figure 1.10). For n-PLS, a regression plot is generated for each column of **Y**. Ideally, the regression plots should have a correlation coefficient close to 1, demonstrating that the property (or properties) of interest is (are) being accurately represented by the chromatographic data. Furthermore, interpretation of the linear regression vectors (LRV) can specify which variables of **X** are positively correlated, negatively correlated, or non-correlated to **Y**. Each LRV can be refolded to produce a plot that visually appears like a comprehensive 2D chromatogram (Figure 1.10). Positive variables in the LRV will correspond to chromatographic variables that are positively correlated with the predicted

variables, whereas anticorrelated variables will be negative. Regions of little or no intensity in the LRVs correspond to chromatographic variables which do not correlate with the predicted property. Outliers in the model can also be detected by examining a plot of the *Q* residuals versus Hotelling's T^2 statistic (Figure 1.10). The *Q* residuals are a measure the difference between the original and modeled data while the Hotelling's T^2 statistic calculates the variation of each sample within the model [120]. Therefore, samples with a high *Q* residual or Hotelling's T^2 could be considered as possible outliers because the samples deviated greatly from the predictions made by the model.

1.4. Overview of the Following Chapters

1.4.1. Chapter 2: Analytical Determination of the Severity of Potato Taste Defect in Roasted East African Arabica Coffee

The quality of East African coffee beans has been significantly reduced by a flavor defect known as potato taste defect (PTD) due to the presence of 2-isopropyl-3-methoxypyrazine (IPMP) and 2-isobutyl-3-methoxypyrazine (IBMP). Therefore, the aims of this study were to determine the correlation between these methoxypyrazines and the severity of odor attributed to PTD and discover additional analytes that may be correlated with PTD using Fisher ratio analysis, a supervised discovery-based data analysis method. Specialty ground roasted coffees from East Africa were classified as clean (i.e., no off-odor), mild, medium, or strong PTD. For the samples examined, IPMP was found to discriminate between non-defective and defective samples, while IBMP did not do so. Samples affected by PTD exhibited a wide range of IPMP concentration (1.6 – 529.9 ng/g). Except for one sample, the IPMP concentration in defective samples was greater than the average IPMP concentration in the non-defective samples (2.0 ng/g). Also, an analysis of variance found that IPMP concentrations were significantly different

based on the severity of odor attributed to PTD ($p < 0.05$). F-ratio analysis discovered 21 additional analytes whose concentrations were statistically different based on the severity of PTD odor ($p < 0.05$). Generally, analytes that were positively correlated with odor severity generally had unpleasant sensory descriptions, while analytes typically associated with desirable aromas were found to be negatively correlated with odor severity. These findings not only show that IPMP concentration can differentiate the severity of PTD but also that changes in the volatile analyte profile of coffee beans induced by PTD can contribute to odor severity.

1.4.2. Chapter 3: Investigating Sensory-Classified Roasted Arabica Coffee with GC×GC-TOFMS and Chemometrics to Understand Potato Taste Defect

The presence of flavor defects in coffee beans can negatively impact quality, the consumer experience, and commercial trade. PTD, a flavor defect specific to East African coffee, is often characterized by a musty, vegetable-like aroma. While previous work has correlated PTD with the presence of IPMP, additional changes in the volatile profile of these beans can further amplify the distinct odor of this defect. The aim of this work was to develop a volatile fingerprint of PTD in roasted arabica coffee using headspace solid-phase microextraction (HS-SPME) coupled to GC×GC-TOFMS and chemometrics. Examination of the HS-SPME-GC×GC-TOFMS data with tile-based F-ratio analysis discovered 359 analytes that differentiated clean coffee samples from those impacted by severe PTD (p-value < 0.01). It was determined that 327 of the identified analytes were more prevalent in the clean coffee samples while 32 analytes, including IPMP, exhibited higher signals in the impacted coffee samples. PCA of the F-ratio results demonstrated that the coffee samples clustered based on the presence of PTD. PLS regression modeling further demonstrated that the compounds discovered by F-ratio analysis were correlated with PTD by accurately predicting the concentration of IPMP in the samples.

Investigation of the compounds highly weighted in both the PCA and PLS loadings suggest that the presence of microorganisms on coffee beans after antestia bug damage could be a potential pathway for PTD. This damage results in an overall decrease of analytes that are known to have positive sensory contributions to coffee aroma. Collectively, the volatile fingerprint shown herein illustrates that PTD alters the biochemical process in coffee beans.

1.4.3. Chapter 4: Detailed Chemical Compositional Analysis of a Thermally Stressed Rocket Fuel using GC×GC-TOFMS and Chemometric Data Analysis

Ensuring reliability, reusability, and operability of Air Force and Space Force propulsion systems motivates the need for quantitative connections between fuel composition, properties, and performance. Predictive computational models in a digital environment are advantageous for correlating accurate fuel property measurements with detailed compositional information of multicomponent fuels like RP-1 and RP-2. To facilitate informed decisions regarding composition, specification, and fit-for-purpose behavior of complex fuels, we apply GC×GC-TOFMS and chemometric data analysis to better understand how fuel performance depends upon chemical composition in the challenging context of fuel thermal stability. In our current investigation, we seek to understand the relationship between the chemical composition of fuels and their application-specific thermal performance properties acquired with the Compact Rapid Assessment of Fuel Thermal Integrity (CRAFTI) experimental platform, providing fuel-specific samples thermally stressed under conditions incorporating real-world engine design features. A multicomponent rocket hydrocarbon fuel, which in its original state is a “clean” fuel (low polars), was subjected to thermal stressing at the following temperatures, ^{o}F: 300, 500, 700, and 900. Chemical composition data for these fuel samples was collected by GC×GC-TOFMS and analyzed using tile-based F-ratio analysis, discovering 92 compounds that changed significantly

in concentration as a function of thermal stress temperature. Most of the 92 compounds were essentially not present in the original sample, and then produced significant signal at 300 ºF. Nearly all these compounds increased in concentration from 300 ºF to 900 ºF, with the largest jump in concentration from 700 ºF to 900 ºF. A variety of compound types were produced: olefins (36), paraffins (33), aromatics (11), and oxygenated compounds (12). This analytical platform has broad implications for the development of high-fidelity composition-property models, leading to an optimized approach to fuel formulation and specification for advanced engine cycles.

1.4.4. Chapter 5: Development of Variance Rank Initiated-Unsupervised Sample Indexing for Gas Chromatography-Mass Spectrometry Analysis

Traditional non-targeted chemometric workflows for GC-MS data rely on using supervised methods, which requires *a priori* knowledge of sample class membership. Herein, we propose a simple, unsupervised chemometric workflow known as variance rank initiated-unsupervised sample indexing (VRI-USI). VRI-USI discovers analyte peaks exhibiting high relative variance across all samples, followed by *k*-means clustering on the individual peaks. Based upon how the samples cluster for a given peak, a sample index assignment is provided. Using a probabilistic argument, if the same sample index assignment appears for several discovered peaks, then this outcome strongly suggests that the samples are properly classified by that particular sample index assignment. Thus, relevant chemical differences between the samples have been discovered in an unsupervised fashion. The VRI-USI workflow is demonstrated on three, increasingly difficult data sets: simulations, yeast metabolomics, and human cancer metabolomics. For simulated GC-MS data sets, VRI-USI discovered 85 – 90% of analytes modeled to vary between sample classes. Nineteen out of 53 peaks in the peak table

developed for the yeast metabolome data set had the same sample index assignments, indicating that those indices are most likely due to class-distinguishing chemical differences. A *t*-test revealed that 22 out of 53 peaks were statistically significant ($p < 0.05$) when using those sample index assignments. Likewise, for the human cancer metabolomics study, VRI-USI discovered 25 analytes that were statistically different ($p < 0.05$) using the sample index assignments determined to highlight meaningful sample-based differences. For all data sets, the sample index assignments that were deduced from VRI-USI were the correct class-based difference when using prior knowledge. VRI-USI holds promise as an exploratory data analysis workflow for studies in which analysts do not readily have *a priori* class information or want to uncover the underlying nature of their data set.

1.4.5. Chapter 6: Enhancing Partial Least Squares Modeling of Comprehensive Two-Dimensional Gas Chromatography Time-of-Flight Mass Spectrometry Data by Tile-Based Variance Ranking

Chemometric methods like PLS regression are valuable for correlating sample-based differences hidden in GC×GC data to independently measured physicochemical properties. Herein, this work establishes the first implementation of tile-based variance ranking as a selective data reduction methodology to improve PLS modeling performance of 58 diverse aerospace fuels. Tile-based variance ranking discovered a total of 521 analytes with a square of the relative standard deviation (RSD^2) in signal between 0.07 to 22.84. The goodness-of-fit for the models were determined by their normalized root-mean-square error of cross-validation (NRMSECV) and normalized root-mean-square error of prediction (NRMSEP). PLS models developed for viscosity, hydrogen content, and heat of combustion using all 521 features discovered by tile-based variance ranking had a respective NRMSECV (NRMSEP) equal to 10.5

% (10.2 %), 8.3 % (7.6 %), and 13.1 % (13.5 %). In contrast, use of a single-grid binning scheme, a common data reduction strategy for PLS analysis, resulted in less accurate models for viscosity (NRMSECV = 14.2 %; NRMSEP = 14.3 %), hydrogen content (NRMSECV = 12.1 %; NRMSEP = 11.0 %), and heat of combustion (NRMSECV = 14.4 %; NRMSEP = 13.6 %). Further, the features discovered by tile-based variance ranking can be optimized for each PLS model with RReliefF analysis, a machine learning algorithm. RReliefF feature optimization selected 48, 125, and 172 analytes out of the original 521 discovered by tile-based variance ranking to model viscosity, hydrogen content, and heat of combustion, respectively. The RReliefF optimized features developed highly accurate property-composition models for viscosity (NRMSECV = 7.9 %; NRMSEP = 5.8 %), hydrogen content (NRMSECV = 7.0 %; NRMSEP = 4.9 %), heat of combustion (NRMSECV = 7.9 %; NRMSEP = 8.4 %). This work also demonstrates that processing the chromatograms with a tile-based approach allows the analyst to directly identify the analytes of importance in a PLS model. Coupling tile-based feature selection with PLS analysis allows for deeper understanding in any property-composition study.

1.4.6. Chapter 7: Tile-Based Pairwise Analysis of GC×GC-TOFMS Data to Facilitate Analyte Discovery and Mass Spectrum Purification

A new tile-based pairwise analysis workflow, termed 1v1 analysis, is presented to discover and identify analytes that differentiate two chromatograms collected using GC×GC-TOFMS. Tile-based 1v1 analysis easily discovered all 18 non-native analytes spiked in diesel fuel within the top 30 hits, outperforming standard pairwise chromatographic analyses. However, eight spiked analytes could not be identified with MCR-ALS nor PARAFAC due to background contamination. Analyte identification was achieved with class comparison enabled-mass

spectrum purification (CCE-MSP), which obtains a pure analyte spectrum by normalizing the spectra to an interferent *m/z* identified from 1v1 analysis and subtracting the two spectra. This report also details the development of CCE-MSP assisted MCR-ALS, which removes the identified interferent *m/z* from the data prior to decomposition. In total, 17 out of 18 spiked analytes had a MV > 800 with both versions of CCE-MSP. For example, MCR-ALS and PARAFAC were unable to decompose the pure spectrum of methyl decanoate (MVs < 200) due to its low $R_{s,2D}$ (~ 0.34) and high interferent-to-analyte signal ratio (~ 30:1). By leveraging information gained from 1v1 analysis, CCE-MSP and CCE-MSP assisted MCR-ALS obtained a pure spectrum with an average MV of 908 and 964, respectively. Furthermore, tile-based 1v1 analysis was applied to track moisture damage in cacao beans, where 86 analytes with at least a 2-fold concentration change were discovered between the unmolded and molded samples. This 1v1 analysis workflow is beneficial for studies where multiple replicates are either unavailable or undesirable to save analysis time.

1.4.7. Chapter 8: Development of an Enhanced Total Ion Current Chromatogram Algorithm to Improve Untargeted Peak Detection

Accurate analyte peak detection from the background noise is a fundamental step in data analysis. Often, this is initially performed on the total ion current chromatogram (TIC), which is the summed signal from all mass spectral channels. Despite the detection of many of the most abundant peaks within a chromatogram, a large fraction of peaks remains undetected in the standard TIC due to their signal being below the limit of detection. To find peaks obscured by background noise, an untargeted peak detection method termed the “enhanced TIC algorithm” was developed for GC×GC-TOFMS. The reported algorithm utilizes the entire mass spectral dimension to find regions of analytical signal above a threshold while zeroing the background

noise. The resulting chromatographic data is summed together to create the enhanced TIC. The utility of the enhanced TIC algorithm is demonstrated using serial dilutions from a 10 parts-per-thousand (ppth) test mixture. For the chromatograms collected at 1 and 10 parts-per-million (ppm), the enhanced TIC algorithm recovered 62 % and 93 %, respectively, of the original peaks observed in the 10 ppth mixture, while the standard TIC recovered only 0 % and 45 %, respectively. The improvement in signal enhancement was also shown on a separation of a yeast cell metabolite extract, where the enhanced TIC found 33 – 64 % more peaks than the standard TIC. Chromatographic simulations with increasing levels of background noise were also conducted to compare the enhanced and standard TICs in the context of SOT. Simulated chromatograms with lower signal-to-noise were more accurately modeled by the SOT after enhanced TIC processing compared to those processed by the standard TIC. The enhanced TIC method demonstrates an immense benefit in peak discovery to improve data analysis efforts.

1.4.8. Chapter 9: Enhancing Gas Chromatography-Mass Spectrometry Resolution and Pure Analyte Discovery using Intra-Chromatogram Elution Profile Matching

The ability to identify and quantify an analyte of interest in GC-MS chromatograms is highly dependent on its R_s and degree of spectral contamination from noise and background interferences. Often, chemometric decomposition methods like MCR-ALS can be employed to mathematically resolve the signal of the analyte from these background interferences to improve identification and quantitation efforts. However, these methods can perform poorly if these chromatographic situations are too challenging. Thus, we propose a novel computational algorithm, termed mzCompare, to improve analyte identification and quantitation when coupled to MCR-ALS. The mzCompare method utilizes an underlying requirement that the retention time and peak shape between *m/z* of the same analyte should be similar. By discovering the selective *m/z* for a given analyte in a chromatogram, a pure elution profile can be generated and used as an

equality constraint in MCR-ALS. The performance of the mzCompare methodology is demonstrated with both experimental and simulated chromatograms. Experimentally, unresolved analytes with a R_s as low as 0.05 could be confidently identified with mzCompare assisted MCR-ALS. Furthermore, application of the mzCompare algorithm to a complex aerospace fuel resulted in the discovery of 335 analytes, a 44 % increase compared to conventional peak detection methods. GC-MS simulations of target-interferent analyte pairs demonstrated that the performance of MCR-ALS deteriorated below a $R_s \sim 0.25$. However, mzCompare assisted MCR-ALS showed excellent identification and acceptable quantitative accuracy at a $R_s \sim 0.02$. Collectively, the results highlighted herein demonstrate how the mzCompare algorithm can help analysts overcome modeling ambiguities resulting from the chemometric multiplex disadvantage.

1.5. References

[1] M.W. Amer, B. Mitrevski, W. Roy Jackson, A.L. Chaffee, P.J. Marriott, Multidimensional and comprehensive two-dimensional gas chromatography of dichloromethane soluble products from a high sulfur Jordanian oil shale, Talanta. 120 (2014) 55–63. https://doi.org/10.1016/j.talanta.2013.11.069.

[2] R.B. Wilson, W.C. Siegler, J.C. Hoggard, B.D. Fitz, J.S. Nadeau, R.E. Synovec, Achieving high peak capacity production for gas chromatography and comprehensive two-dimensional gas chromatography by minimizing off-column peak broadening, J. Chromatogr. A. 1218 (2011) 3130–3139. https://doi.org/10.1016/j.chroma.2010.12.108.

[3] J. Krupčík, P. Májek, R. Gorovenko, I. Špánik, P. Sandra, D.W. Armstrong, On the determination of a detector response enhancement factor for flow modulated comprehensive two-dimensional gas chromatography, J. Chromatogr. A. 1286 (2013) 235–240. https://doi.org/10.1016/j.chroma.2013.02.068.

[4] P. Korytár, P.E.G. Leonards, J. De Boer, U.A.T. Brinkman, Group separation of organohalogenated compounds by means of comprehensive two-dimensional gas chromatography, J. Chromatogr. A. 1086 (2005) 29–44. https://doi.org/10.1016/j.chroma.2005.05.087.

[5] P. Haglund, P. Korytár, C. Danielsson, J. Diaz, K. Wiberg, P. Leonards, U.A.T. Brinkman, J. De Boer, GC×GC-ECD: A promising method for the determination of dioxins and dioxin-like PCBs in food and feed, Anal. Bioanal. Chem. 390 (2008) 1815–1827. https://doi.org/10.1007/s00216-008-1896-0.

[6] K. Murtada, D. Bowman, M. Edwards, J. Pawliszyn, Thin-film microextraction combined with comprehensive two-dimensional gas chromatography time-of-flight mass spectrometry screening for presence of multiclass organic pollutants in drinking water samples, Talanta. 242 (2022) 123301. https://doi.org/10.1016/j.talanta.2022.123301.

[7] P.H. Stefanuto, K.A. Perrault, L.M. Dubois, B. L'Homme, C. Allen, C. Loughnane, N. Ochiai, J.F. Focant, Advanced method optimization for volatile aroma profiling of beer using two-dimensional gas chromatography time-of-flight mass spectrometry, J. Chromatogr. A. 1507 (2017) 45–52. https://doi.org/10.1016/j.chroma.2017.05.064.

[8] G. Ebadzadsahrai, E.A. Higgins Keppler, S.D. Soby, H.D. Bean, Inhibition of Fungal Growth and Induction of a Novel Volatilome in Response to Chromobacterium vaccinii Volatile Organic Compounds, Front. Microbiol. 11 (2020). https://doi.org/10.3389/fmicb.2020.01035.

[9] J. Omar, M. Olivares, J.M. Amigo, N. Etxebarria, Resolution of co-eluting compounds of Cannabis Sativa in comprehensive two-dimensional gas chromatography/mass spectrometry detection with Multivariate Curve Resolution-Alternating Least Squares, Talanta. 121 (2014) 273–280. https://doi.org/10.1016/j.talanta.2013.12.044.

[10] L.W. Hantao, B.R. Toledo, F.A. De Lima Ribeiro, M. Pizetta, C.G. Pierozzi, E.L. Furtado, F. Augusto, Comprehensive two-dimensional gas chromatography combined to multivariate data analysis for detection of disease-resistant clones of Eucalyptus, Talanta. 116 (2013) 1079–1084. https://doi.org/10.1016/j.talanta.2013.08.033.

[11] J. Krupčík, R. Gorovenko, I. Špánik, P. Sandra, D.W. Armstrong, Flow-modulated comprehensive two-dimensional gas chromatography with simultaneous flame ionization and quadrupole mass spectrometric detection, J. Chromatogr. A. 1280 (2013) 104–111. https://doi.org/10.1016/j.chroma.2013.01.015.

[12] J.M. Davis, J.C. Giddings, Statistical theory of component overlap in multicomponent chromatograms, Anal. Chem. 55 (1983) 418–424. https://doi.org/10.1021/ac00254a003.

[13] P.Q. Tranchida, D. Sciarrone, P. Dugo, L. Mondello, Heart-cutting multidimensional gas chromatography: A review of recent evolution, applications, and future prospects, Anal. Chim. Acta. 716 (2012) 66–75. https://doi.org/10.1016/J.ACA.2011.12.015.

[14] Z. Liu, J.B. Phillips, Comprehensive Two-Dimensional Gas Chromatography using an On-Column Thermal Modulator Interface, J. Chromatogr. Sci. 29 (1991) 227–231. https://doi.org/10.1093/chromsci/29.6.227.

[15] R.E. Murphy, M.R. Schure, J.P. Foley, Effect of Sampling Rate on Resolution in Comprehensive Two-Dimensional Liquid Chromatography, Anal. Chem. 70 (1998) 1585–1594. https://doi.org/10.1021/ac971184b.

[16] J. V. Seeley, Theoretical study of incomplete sampling of the first dimension in comprehensive two-dimensional chromatography, J. Chromatogr. A. 962 (2002) 21–27. https://doi.org/10.1016/S0021-9673(02)00461-2.

[17] O. Panić, T. Górecki, C. McNeish, A.H. Goldstein, B.J. Williams, D.R. Worton, S. V. Hering, N.M. Kreisberg, Development of a new consumable-free thermal modulator for comprehensive two-dimensional gas chromatography, J. Chromatogr. A. 1218 (2011) 3070–3079. https://doi.org/10.1016/j.chroma.2011.03.024.

[18] S.J. Kim, G. Serrano, K.D. Wise, K. Kurabayashi, E.T. Zellers, Evaluation of a microfabricated thermal modulator for comprehensive two-dimensional microscale gas chromatography, Anal. Chem. 83 (2011) 5556–5562. https://doi.org/10.1021/ac200336e.

[19] J. Luong, X. Guan, S. Xu, R. Gras, R.A. Shellie, Thermal Independent Modulator for Comprehensive Two-Dimensional Gas Chromatography, Anal. Chem. 88 (2016) 8428–8432. https://doi.org/10.1021/acs.analchem.6b02525.

[20] A.M. Muscalu, M. Edwards, T. Górecki, E.J. Reiner, Evaluation of a single-stage consumable-free modulator for comprehensive two-dimensional gas chromatography: Analysis of polychlorinated biphenyls, organochlorine pesticides and chlorobenzenes, J. Chromatogr. A. 1391 (2015) 93–101. https://doi.org/10.1016/j.chroma.2015.02.074.

[21] B. Savareear, M. Brokl, C. Wright, J.F. Focant, Thermal desorption comprehensive two-dimensional gas chromatography coupled to time of flight mass spectrometry for vapour phase mainstream tobacco smoke analysis, J. Chromatogr. A. 1525 (2017) 126–137. https://doi.org/10.1016/j.chroma.2017.10.013.

[22] B. V. Burger, T. Snyman, W.J.G. Burger, W.F. Van Rooyen, Thermal modulator array for analyte modulation and comprehensive two-dimensional gas chromatography, J. Sep. Sci. 26 (2003) 123–128. https://doi.org/10.1002/jssc.200390002.

[23] A. Ghosh, C.T. Bates, S.K. Seeley, J. V. Seeley, High speed Deans switch for low duty cycle comprehensive two-dimensional gas chromatography, J. Chromatogr. A. 1291 (2013) 146–154. https://doi.org/10.1016/j.chroma.2013.04.003.

[24] J. V. Seeley, N.E. Schimmel, S.K. Seeley, The multi-mode modulator: A versatile fluidic device for two-dimensional gas chromatography, J. Chromatogr. A. 1536 (2018) 6–15. https://doi.org/10.1016/j.chroma.2017.06.030.

[25] P.A. Bueno, J. V. Seeley, Flow-switching device for comprehensive two-dimensional gas chromatography, J. Chromatogr. A. 1027 (2004) 3–10. https://doi.org/10.1016/j.chroma.2003.10.033.

[26] J. V. Seeley, N.J. Micyus, S. V. Bandurski, S.K. Seeley, J.D. McCurry, Microfluidic deans switch for comprehensive two-dimensional gas chromatography, Anal. Chem. 79 (2007) 1840–1847. https://doi.org/10.1021/ac061881g.

[27] C.E. Freye, R.E. Synovec, High temperature diaphragm valve-based comprehensive two-dimensional gas chromatography with time-of-flight mass spectrometry, Talanta. 161 (2016) 675–680. https://doi.org/10.1016/j.talanta.2016.09.002.

[28] R.E. Mohler, B.J. Prazen, R.E. Synovec, Total-transfer, valve-based comprehensive two-dimensional gas chromatography, Anal. Chim. Acta. 555 (2006) 68–74. https://doi.org/10.1016/j.aca.2005.08.072.

[29] P.Q. Tranchida, G. Purcaro, A. Visco, L. Conte, P. Dugo, P. Dawes, L. Mondello, A flexible loop-type flow modulator for comprehensive two-dimensional gas chromatography, J. Chromatogr. A. 1218 (2011) 3140–3145. https://doi.org/10.1016/j.chroma.2010.11.082.

[30] P.Q. Tranchida, S. Salivo, F.A. Franchina, L. Mondello, Flow-modulated comprehensive two-dimensional gas chromatography combined with a high-resolution time-of-flight

mass spectrometer: A proof-of-principle study, Anal. Chem. 87 (2015) 2925–2930. https://doi.org/10.1021/ac5044175.

[31] T.J. Trinklein, S. Schöneich, P.E. Sudol, C.G. Warren, D. V. Gough, R.E. Synovec, Total-transfer comprehensive three-dimensional gas chromatography with time-of-flight mass spectrometry, J. Chromatogr. A. 1634 (2020) 461654. https://doi.org/10.1016/j.chroma.2020.461654.

[32] S. Schöneich, D. V Gough, T.J. Trinklein, R.E. Synovec, Dynamic pressure gradient modulation for comprehensive two-dimensional gas chromatography with time-of-flight mass spectrometry detection, J. Chromatogr. A. (2020) 460982. https://doi.org/https://doi.org/10.1016/j.chroma.2020.460982.

[33] P.Q. Tranchida, F.A. Franchina, P. Dugo, L. Mondello, Flow-modulation low-pressure comprehensive two-dimensional gas chromatography, J. Chromatogr. A. 1372 (2014) 236–244.

[34] P.Q. Tranchida, F.A. Franchina, P. Dugo, L. Mondello, Use of greatly-reduced gas flows in flow-modulated comprehensive two-dimensional gas chromatography-mass spectrometry, J. Chromatogr. A. 1359 (2014) 271–276.

[35] F.A. Franchina, M. Maimone, P.Q. Tranchida, L. Mondello, Flow modulation comprehensive two-dimensional gas chromatography-mass spectrometry using ≈ 4 mL min-1 gas flows, J. Chromatogr. A. 1441 (2016) 134–139. https://doi.org/10.1016/j.chroma.2016.02.041.

[36] P.Q. Tranchida, F.A. Franchina, M. Zoccali, I. Bonaccorsi, F. Cacciola, L. Mondello, A direct sensitivity comparison between flow-modulated comprehensive 2D and 1D GC in untargeted and targeted MS-based experiments, J. Sep. Sci. 36 (2013) 2746–2752. https://doi.org/10.1002/jssc.201300423.

[37] T.J. Trinklein, D. V. Gough, C.G. Warren, G.S. Ochoa, R.E. Synovec, Dynamic pressure gradient modulation for comprehensive two-dimensional gas chromatography, J. Chromatogr. A. 1609 (2020) 460488. https://doi.org/10.1016/j.chroma.2019.460488.

[38] S. Schöneich, T.J. Trinklein, C.G. Warren, R.E. Synovec, A systematic investigation of comprehensive two-dimensional gas chromatography time-of-flight mass spectrometry with dynamic pressure gradient modulation for high peak capacity separations, Anal. Chim. Acta. 1134 (2020) 115–124. https://doi.org/10.1016/j.aca.2020.08.023.

[39] P.E. Sudol, M. Galletta, P.Q. Tranchida, M. Zoccali, L. Mondello, R.E. Synovec, Untargeted profiling and differentiation of geographical variants of wine samples using headspace solid-phase microextraction flow-modulated comprehensive two-dimensional gas chromatography with the support of tile-based Fisher ratio analysis, J. Chromatogr. A. 1662 (2022) 462735. https://doi.org/10.1016/j.chroma.2021.462735.

[40] P.M. Harvey, R.A. Shellie, Factors affecting peak shape in comprehensive two-dimensional gas chromatography with non-focusing modulation, J. Chromatogr. A. 1218 (2011) 3153–3158. https://doi.org/10.1016/j.chroma.2010.08.029.

[41] W.C. Siegler, B.D. Fitz, J.C. Hoggard, R.E. Synovec, Experimental study of the quantitative precision for valve-based comprehensive two-dimensional gas chromatography, Anal. Chem. 83 (2011) 5190–5196. https://doi.org/10.1021/ac200302b.

[42] Z. Liu, D.G. Patterson, M.L. Lee, Geometric approach to factor analysis for the estimation of orthogonality and practical peak capacity in comprehensive two-dimensional separations, Anal. Chem. 67 (1995) 3840–3845. https://doi.org/10.1021/ac00117a004.

[43] A.L. Lee, K.D. Bartle, A.C. Lewis, A model of peak amplitude enhancement in orthogonal two-dimensional gas chromatography, Anal. Chem. 73 (2001) 1330–1335. https://doi.org/10.1021/ac001120s.

[44] M.S. Klee, J. Cochran, M. Merrick, L.M. Blumberg, Evaluation of conditions of comprehensive two-dimensional gas chromatography that yield a near-theoretical maximum in peak capacity gain, J. Chromatogr. A. 1383 (2015) 151–159. https://doi.org/10.1016/j.chroma.2015.01.031.

[45] C.E. Freye, N.R. Moore, R.E. Synovec, Enhancing the chemical selectivity in discovery-based analysis with tandem ionization time-of-flight mass spectrometry detection for comprehensive two-dimensional gas chromatography, J. Chromatogr. A. 1537 (2018) 99–108. https://doi.org/10.1016/j.chroma.2018.01.008.

[46] B. Giocastro, M. Zoccali, P.Q. Tranchida, L. Mondello, Evaluation of different internal diameter coated modulation columns within the context of solid-state modulation, J. Sep. Sci. (2021) 1–20. https://doi.org/10.1002/jssc.202001031.

[47] J. Crucello, L.F.O. Miron, V.H.C. Ferreira, H. Nan, M.O.M. Marques, P.S. Ritschel, M.C. Zanus, J.L. Anderson, R.J. Poppi, L.W. Hantao, Characterization of the aroma profile of novel Brazilian wines by solid-phase microextraction using polymeric ionic liquid sorbent coatings, Anal. Bioanal. Chem. 410 (2018) 4749–4762. https://doi.org/10.1007/s00216-018-1134-3.

[48] F. Stilo, C. Bicchi, A. Robbat, S.E. Reichenbach, C. Cordero, Untargeted approaches in food-omics: The potential of comprehensive two-dimensional gas chromatography/mass spectrometry, TrAC - Trends Anal. Chem. 135 (2021) 116162. https://doi.org/10.1016/j.trac.2020.116162.

[49] V.G. Van Mispelaar, A.C. Tas, A.K. Smilde, P.J. Schoenmakers, A.C. Van Asten, Quantitative analysis of target components by comprehensive two-dimensional gas chromatography, J. Chromatogr. A. 1019 (2003) 15–29. https://doi.org/10.1016/j.chroma.2003.08.101.

[50] J.F. Griffith, W.L. Winniford, K. Sun, R. Edam, J.C. Luong, A reversed-flow differential flow modulator for comprehensive two-dimensional gas chromatography, J. Chromatogr. A. 1226 (2012) 116–123. https://doi.org/10.1016/j.chroma.2011.11.036.

[51] B. Gruber, B.A. Weggler, R. Jaramillo, K.A. Murrell, P.K. Piotrowski, F.L. Dorman, Comprehensive two-dimensional gas chromatography in forensic science: A critical review of recent trends, TrAC - Trends Anal. Chem. 105 (2018) 292–301. https://doi.org/10.1016/j.trac.2018.05.017.

[52] P.E. Sudol, K.M. Pierce, S.E. Prebihalo, K.J. Skogerboe, B.W. Wright, R.E. Synovec, Development of gas chromatographic pattern recognition and classification tools for compliance and forensic analyses of fuels: A review, Anal. Chim. Acta. 1132 (2020) 157-186. https://doi.org/10.1016/j.aca.2020.07.027.

[53] H.D. Bean, J.E. Hill, J.M.D. Dimandja, Improving the quality of biomarker candidates in untargeted metabolomics via peak table-based alignment of comprehensive two-dimensional gas chromatography-mass spectrometry data, J. Chromatogr. A. 1394 (2015) 111–117. https://doi.org/10.1016/j.chroma.2015.03.001.

[54] Y. Zushi, J. Gros, Q. Tao, S.E. Reichenbach, S. Hashimoto, J.S. Arey, Pixel-by-pixel correction of retention time shifts in chromatograms from comprehensive two-dimensional gas chromatography coupled to high resolution time-of-flight mass spectrometry, J. Chromatogr. A. 1508 (2017) 121–129. https://doi.org/10.1016/j.chroma.2017.05.065.

[55] K. Ralston-Hooper, A. Hopf, C. Oh, X. Zhang, J. Adamec, M.S. Sepúlveda, Development of GCxGC/TOF-MS metabolomics for use in ecotoxicological studies with invertebrates, Aquat. Toxicol. 88 (2008) 48–52. https://doi.org/10.1016/j.aquatox.2008.03.002.

[56] E. Hoh, N.G. Dodder, S.J. Lehotay, K.C. Pangallo, C.M. Reddy, K.A. Maruya, Nontargeted comprehensive two-dimensional gas chromatography/time-of-flight mass spectrometry method and software for inventorying persistent and bioaccumulative contaminants in marine environments, Environ. Sci. Technol. 46 (2012) 8001–8008. https://doi.org/10.1021/es301139q.

[57] D. Zanella, A. Henin, S. Mascrez, P.H. Stefanuto, F.A. Franchina, J.F. Focant, G. Purcaro, Comprehensive two-dimensional gas chromatographic platforms comparison for exhaled breath metabolites analysis, J. Sep. Sci. 45 (2022) 3542–3555. https://doi.org/10.1002/jssc.202200164.

[58] K.M. Pierce, J.L. Hope, K.J. Johnson, B.W. Wright, R.E. Synovec, Classification of gasoline data obtained by gas chromatography using a piecewise alignment algorithm combined with feature selection and principal component analysis, J. Chromatogr. A. 1096 (2005) 101–110. https://doi.org/10.1016/j.chroma.2005.04.078.

[59] D. Zhang, X. Huang, F.E. Regnier, M. Zhang, Two-dimensional correlation optimized warping algorithm for aligning GCxGC-MS data, Anal. Chem. 80 (2008) 2664–2671. https://doi.org/10.1021/ac7024317.

[60] J. Vial, H. Noçairi, P. Sassiat, S. Mallipatu, G. Cognon, D. Thiébaut, B. Teillet, D.N. Rutledge, Combination of dynamic time warping and multivariate analysis for the comparison of comprehensive two-dimensional gas chromatograms. Application to plant extracts, J. Chromatogr. A. 1216 (2009) 2866–2872. https://doi.org/10.1016/j.chroma.2008.09.027.

[61] B. Wang, A. Fang, J. Heim, B. Bogdanov, S. Pugh, M. Libardoni, X. Zhang, DISCO: Distance and spectrum correlation optimization alignment for two-dimensional gas chromatography time-of-flight mass spectrometry-based metabolomics, Anal. Chem. 82 (2010) 5069–5081. https://doi.org/10.1021/ac100064b.

[62] R.C. Allen, S.C. Rutan, Semi-automated alignment and quantification of peaks using parallel factor analysis for comprehensive two-dimensional liquid chromatography-diode array detector data sets, Anal. Chim. Acta. 723 (2012) 7–17. https://doi.org/10.1016/j.aca.2012.02.019.

[63] P.E. Sudol, D. V. Gough, S.E. Prebihalo, R.E. Synovec, Impact of data bin size on the classification of diesel fuels using comprehensive two-dimensional gas chromatography with principal component analysis, Talanta. 206 (2020) 120239. https://doi.org/10.1016/j.talanta.2019.120239.

[64] S. Furbo, A.B. Hansen, T. Skov, J.H. Christensen, Pixel-based analysis of comprehensive two-dimensional gas chromatograms (color plots) of petroleum: A tutorial, Anal. Chem. 86 (2014) 7160–7170. https://doi.org/10.1021/ac403650d.

[65] J. Kiefl, C. Cordero, L. Nicolotti, P. Schieberle, S.E. Reichenbach, C. Bicchi, Performance evaluation of non-targeted peak-based cross-sample analysis for comprehensive two-dimensional gas chromatography-mass spectrometry data and application to processed hazelnut profiling, J. Chromatogr. A. 1243 (2012) 81–90. https://doi.org/10.1016/j.chroma.2012.04.048.

[66] T.J. Davis, T.R. Firzli, E.A. Higgins Keppler, M. Richardson, H.D. Bean, Addressing Missing Data in GC × GC Metabolomics: Identifying Missingness Type and Evaluating the Impact of Imputation Methods on Experimental Replication, Anal. Chem. 94 (2022) 10912–10920. https://doi.org/10.1021/acs.analchem.1c04093.

[67] L.C. Marney, W. Christopher Siegler, B.A. Parsons, J.C. Hoggard, B.W. Wright, R.E. Synovec, Tile-based Fisher-ratio software for improved feature selection analysis of comprehensive two-dimensional gas chromatography-time-of-flight mass spectrometry data, Talanta. 115 (2013) 887–895. https://doi.org/10.1016/j.talanta.2013.06.038.

[68] B.A. Parsons, L.C. Marney, W.C. Siegler, J.C. Hoggard, B.W. Wright, R.E. Synovec, Tile-Based Fisher Ratio Analysis of Comprehensive Two-Dimensional Gas Chromatography Time-of-Flight Mass Spectrometry (GC × GC–TOFMS) Data Using a Null Distribution Approach, Anal. Chem. 87 (2015) 3812–3819. https://doi.org/10.1021/ac504472s.

[69] T.J. Trinklein, R.E. Synovec, Simulating comprehensive two-dimensional gas chromatography mass spectrometry data with realistic run-to-run shifting to evaluate the robustness of tile-based Fisher ratio analysis, J. Chromatogr. A. 1677 (2022) 463321. https://doi.org/10.1016/j.chroma.2022.463321.

[70] S. Schöneich, C.N. Cain, C.E. Freye, R.E. Synovec, Optimization of Parameters for ROI Data Compression for Nontargeted Analyses Using LC–HRMS, Anal. Chem. 95 (2022) 1513–1521. https://doi.org/10.1021/acs.analchem.2c04538.

[71] S. Schöneich, C.N. Cain, P.E. Sudol, R.E. Synovec, Enabling cuboid-based fisher ratio analysis using total-transfer comprehensive three-dimensional gas chromatography with time-of-flight mass spectrometry, J. Chromatogr. A. 1708 (2023) 464341. https://doi.org/10.1016/j.chroma.2023.464341.

[72] O. Amador-Muñoz, P.J. Marriott, Quantification in comprehensive two-dimensional gas chromatography and a model of quantification based on selected summed modulated peaks, J. Chromatogr. A. 1184 (2008) 323–340. https://doi.org/10.1016/j.chroma.2007.10.041.

[73] D.W. Cook, M.L. Burnham, D.C. Harmes, D.R. Stoll, S.C. Rutan, Comparison of multivariate curve resolution strategies in quantitative LCxLC: Application to the

quantification of furanocoumarins in apiaceous vegetables, Anal. Chim. Acta. 961 (2017) 49–58. https://doi.org/10.1016/j.aca.2017.01.047.

[74] R. Tauler, Multivariate curve resolution applied to second order data, Chemom. Intell. Lab. Syst. 30 (1995) 133–146. https://doi.org/10.1016/0169-7439(95)00047-X.

[75] S.C. Rutan, A. de Juan, R. Tauler, Introduction to Multivariate Curve Resolution, in: S.D. Brown, R. Tauler, B. Walczak (Eds.), Compr. Chemom., Vol. 2, Elsevier, 2009: pp. 249–259.

[76] A. de Juan, J. Jaumot, R. Tauler, Multivariate curve resolution (MCR): Solving the mixture analysis problem, Anal. Methods. 6 (2014) 4964–4976. https://doi.org/10.1039/C4AY00571F.

[77] H. Parastar, J.R. Radović, M. Jalali-Heravi, S. Diez, J.M. Bayona, R. Tauler, Resolution and quantification of complex mixtures of polycyclic aromatic hydrocarbons in heavy fuel oil sample by means of GC × GC-TOFMS combined to multivariate curve resolution, Anal. Chem. 83 (2011) 9289–9297. https://doi.org/10.1021/ac201799r.

[78] R. Tauler, Calculation of maximum and minimum band boundaries of feasible solutions for species profiles obtained by multivariate curve resolution, J. Chemom. 15 (2001) 627–646. https://doi.org/10.1002/cem.654.

[79] J. Jaumot, R. Tauler, MCR-BANDS: A user friendly MATLAB program for the evaluation of rotation ambiguities in Multivariate Curve Resolution, Chemom. Intell. Lab. Syst. 103 (2010) 96–107. https://doi.org/10.1016/j.chemolab.2010.05.020.

[80] W. Windig, J. Guilment, Interactive Self-Modeling Mixture Analysis, Anal. Chem. 63 (1991) 1425–1432.

[81] F. Cuesta Sánchez, J. Toft, B. Van den Bogaert, D.L. Massart, Orthogonal projection approach applied to peak purity assessment, Anal. Chem. 68 (1996) 79–85. https://doi.org/10.1021/ac950496g.

[82] E.R. Malinowski, Obtaining the key set of typical vectors by factor analysis and subsequent isolation of component spectra, Anal. Chim. Acta. 134 (1982) 129–137. https://doi.org/10.1016/S0003-2670(01)84184-2.

[83] K.J. Schostack, E.R. Malinowski, Investigation of window factor analysis and matrix regression analysis in chromatography, Chemom. Intell. Lab. Syst. 20 (1993) 173–182. https://doi.org/10.1016/0169-7439(93)80013-8.

[84] H.P. Bailey, S.C. Rutan, Chemometric resolution and quantification of four-way data arising from comprehensive 2D-LC-DAD analysis of human urine, Chemom. Intell. Lab. Syst. 106 (2011) 131–141. https://doi.org/10.1016/j.chemolab.2010.07.008.

[85] D.W. Cook, S.C. Rutan, D.R. Stoll, P.W. Carr, Two dimensional assisted liquid chromatography - a chemometric approach to improve accuracy and precision of quantitation in liquid chromatography using 2D separation, dual detectors, and multivariate curve resolution, Anal. Chim. Acta. 859 (2015) 87–95. https://doi.org/10.1016/j.aca.2014.12.009.

[86] R. Bro, PARAFAC. Tutorial and applications, Chemom. Intell. Lab. Syst. 38 (1997) 149–171. https://doi.org/10.1016/S0169-7439(97)00032-4.

[87] D.K. Pinkerton, B.A. Parsons, T.J. Anderson, R.E. Synovec, Trilinearity deviation ratio: A new metric for chemometric analysis of comprehensive two-dimensional gas chromatography time-of-flight mass spectrometry data, Anal. Chim. Acta. 871 (2015) 66–76. https://doi.org/10.1016/j.aca.2015.02.040.

[88] J.C. Hoggard, R.E. Synovec, Parallel factor analysis (PARAFAC) of target analytes in GC × GC-TOFMS data: Automated selection of a model with an appropriate number of factors, Anal. Chem. 79 (2007) 1611–1619. https://doi.org/10.1021/ac061710b.

[89] R.A. Harshman, M.E. Lundy, PARAFAC: Parallel factor analysis, Comput. Stat. Data Anal. 18 (1994) 39–72. https://doi.org/10.1016/0167-9473(94)90132-5.

[90] R.C. Allen, S.C. Rutan, Investigation of interpolation techniques for the reconstruction of the first dimension of comprehensive two-dimensional liquid chromatography-diode array detector data, Anal. Chim. Acta. 705 (2011) 253–260. https://doi.org/10.1016/j.aca.2011.06.022.

[91] Y. Izadmanesh, E. Garreta-Lara, J.B. Ghasemi, S. Lacorte, V. Matamoros, R. Tauler, Chemometric analysis of comprehensive two dimensional gas chromatography–mass spectrometry metabolomics data, J. Chromatogr. A. 1488 (2017) 113–125. https://doi.org/10.1016/j.chroma.2017.01.052.

[92] S. Schöneich, D. V. Gough, T.J. Trinklein, R.E. Synovec, Dynamic pressure gradient modulation for comprehensive two-dimensional gas chromatography with time-of-flight mass spectrometry detection, J. Chromatogr. A. 1620 (2020) 460982. https://doi.org/10.1016/j.chroma.2020.460982.

[93] S.E. Prebihalo, D.K. Pinkerton, R.E. Synovec, Impact of comprehensive two-dimensional gas chromatography time-of-flight mass spectrometry experimental design on data trilinearity and parallel factor analysis deconvolution, J. Chromatogr. A. 1605 (2019) 460368. https://doi.org/10.1016/j.chroma.2019.460368.

[94] T. Skov, J.C. Hoggard, R. Bro, R.E. Synovec, Handling within run retention time shifts in two-dimensional chromatography data using shift correction and modeling, J. Chromatogr. A. 1216 (2009) 4020–4029. https://doi.org/10.1016/j.chroma.2009.02.049.

[95] J.M. Amigo, T. Skov, R. Bro, J. Coello, S. Maspoch, Solving GC-MS problems with PARAFAC2, TrAC - Trends Anal. Chem. 27 (2008) 714–725. https://doi.org/10.1016/j.trac.2008.05.011.

[96] S. Wold, K. Esbensen, P. Geladi, Principal component analysis, Chemom. Intell. Lab. Syst. 2 (1987) 37–52. https://doi.org/10.1016/0169-7439(87)80084-9.

[97] D.A. Jackson, Stopping Rules in Principal Components Analysis: A Comparison of Heuristical and Statistical Approaches, Ecology. 74 (1993) 2204–2214. https://doi.org/10.2307/1939574.

[98] R. Henrion, N-way principal component analysis theory, algorithms and applications, Chemom. Intell. Lab. Syst. 25 (1994) 1–23. https://doi.org/10.1016/0169-7439(93)E0086-J.

[99] R. De Maesschalck, D. Jouan-Rimbaud, D.L. Massart, The Mahalanobis distance, Chemom. Intell. Lab. Syst. 50 (2000) 1–18. https://doi.org/10.1016/S0169-7439(99)00047-7.

[100] N.A. Sinkov, J.J. Harynuk, Cluster resolution: A metric for automated, objective and optimized feature selection in chemometric modeling, Talanta. 83 (2011) 1079–1087. https://doi.org/10.1016/j.talanta.2010.10.025.

[101] B. Worley, S. Halouska, R. Powers, Utilities for quantifying separation in PCA/PLS-DA scores plots, Anal. Biochem. 433 (2013) 102–104. https://doi.org/10.1016/j.ab.2012.10.011.

[102] G. Malmquist, R. Danielsson, Alignment of chromatographic profiles for principal component analysis: a prerequisite for fingerprinting methods, J. Chromatogr. A. 687 (1994) 71–88. https://doi.org/10.1016/0021-9673(94)00726-8.

[103] J.B. MacQueen, Some Methods for Classification and Analysis of Multivariate Observations, in: Proc. Fifth Berkeley Symp. Math. Stat. Probab., 1967: pp. 281–297.

[104] A.K. Jain, Data clustering: 50 years beyond K-means, Pattern Recognit. Lett. 31 (2010) 651–666. https://doi.org/10.1016/j.patrec.2009.09.011.

[105] D. Arthur, S. Vassilvitskii, k-means++: The advantages of careful seeding, in: Proc. Eighteenth Annu. ACM-SIAM Symp. Discret. Algorithms, Society for Industrial and Applied Mathematics, Philadelphia, PA, 2007: pp. 1027–1035.

[106] M. Charrad, N. Ghazzali, V. Boiteau, A. Niknafs, Nbclust: An R package for determining the relevant number of clusters in a data set, J. Stat. Softw. 61 (2014) 1–36. https://doi.org/10.18637/jss.v061.i06.

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

[108] D.L. Davies, D.W. Bouldin, A Cluster Separation Measure, IEEE Trans. Pattern Anal. Mach. Intell. PAMI-1 (1979) 224–227. https://doi.org/10.1109/TPAMI.1979.4766909.

[109] S.E. Prebihalo, G.S. Ochoa, K.L. Berrier, K.J. Skogerboe, K.L. Cameron, J.R. Trump, S.J. Svoboda, J.K. Wickiser, R.E. Synovec, Control-Normalized Fisher Ratio Analysis of Comprehensive Two-Dimensional Gas Chromatography Time-of-Flight Mass Spectrometry Data for Enhanced Biomarker Discovery in a Metabolomic Study of Orthopedic Knee-Ligament Injury, Anal. Chem. 92 (2020) 15526–15533. https://doi.org/10.1021/acs.analchem.0c03456.

[110] C. Brownie, D.D. Boos, J. Hughes-oliver, Modifying the t and ANOVA F Tests When Treatment Is Expected to Increase Variability Relative to Controls, Biometrics. 46 (2019) 259–266.

[111] S. Schöneich, G.S. Ochoa, C.M. Monzón, R.E. Synovec, Minimum variance optimized Fisher ratio analysis of comprehensive two-dimensional gas chromatography / mass spectrometry data: Study of the pacu fish metabolome, J. Chromatogr. A. 1667 (2022) 462868. https://doi.org/10.1016/j.chroma.2022.462868.

[112] K.J. Johnson, R.E. Synovec, Pattern recognition of jet fuels: Comprehensive GC × GC with ANOVA-based feature selection and principal component analysis, Chemom. Intell. Lab. Syst. 60 (2002) 225–237. https://doi.org/10.1016/S0169-7439(01)00198-8.

[113] B.A. Parsons, D.K. Pinkerton, B.W. Wright, R.E. Synovec, Chemical characterization of the acid alteration of diesel fuel: Non-targeted analysis by two-dimensional gas chromatography coupled with time-of-flight mass spectrometry with tile-based Fisher ratio and combinatorial threshold determination, J. Chromatogr. A. 1440 (2016) 179–190. https://doi.org/10.1016/j.chroma.2016.02.067.

[114] G.S. Ochoa, S.E. Prebihalo, B.C. Reaser, L.C. Marney, R.E. Synovec, Statistical inference of mass channel purity from Fisher ratio analysis using comprehensive two-dimensional gas chromatography with time of flight mass spectrometry data, J. Chromatogr. A. 1627 (2020) 461401. https://doi.org/10.1016/j.chroma.2020.461401.

[115] S. Wold, M. Sjöström, L. Eriksson, PLS-regression: A basic tool of chemometrics, Chemom. Intell. Lab. Syst. 58 (2001) 109–130. https://doi.org/10.1016/S0169-7439(01)00155-1.

[116] B. Kehimkar, J.C. Hoggard, L.C. Marney, M.C. Billingsley, C.G. Fraga, T.J. Bruno, R.E. Synovec, Correlation of rocket propulsion fuel properties with chemical composition using comprehensive two-dimensional gas chromatography with time-of-flight mass spectrometry followed by partial least squares regression analysis, J. Chromatogr. A. 1327 (2014) 132–140. https://doi.org/10.1016/j.chroma.2013.12.060.

[117] K.L. Berrier, C.E. Freye, M.C. Billingsley, R.E. Synovec, Predictive Modeling of Aerospace Fuel Properties Using Comprehensive Two-Dimensional Gas Chromatography with Time-Of-Flight Mass Spectrometry and Partial Least Squares Analysis, Energy Fuels. 34 (2020) 4084–4094. https://doi.org/10.1021/acs.energyfuels.9b04108.

[118] V. Abrahamsson, N. Ristic, K. Franz, K. Van Geem, Comprehensive two-dimensional gas chromatography in combination with pixel-based analysis for fouling tendency prediction, J. Chromatogr. A. 1501 (2017) 89–98. https://doi.org/10.1016/j.chroma.2017.04.021.

[119] V. Consonni, G. Baccolo, F. Gosetti, R. Todeschini, D. Ballabio, A MATLAB toolbox for multivariate regression coupled with variable selection, Chemom. Intell. Lab. Syst. 213 (2021) 104313. https://doi.org/10.1016/j.chemolab.2021.104313.

[120] D. Ballabio, V. Consonni, Classification tools in chemistry. Part 1: Linear models. PLS-DA, Anal. Methods. 5 (2013) 3790–3798. https://doi.org/10.1039/c3ay40582f.

Chapter 2: Analytical Determination of the Severity of Potato Taste Defect in Roasted East African Arabica Coffee

2.1. Introduction

As coffee consumption increases globally, demands for specialty and/or high-quality coffee beans are primarily driven by consumer preferences towards the taste and aroma of the final brew. The sensory properties desired by consumers are due to biochemical reactions that occur within the coffee bean as it undergoes various pre- (i.e., genetics, geographic origin, and environmental conditions) and post-harvesting (i.e., processing, roasting, and preparation) processes [1,2]. Since aroma is the most important evaluation criteria for coffee bean quality, the presence of defects, such as under-ripened, over-ripened or insect damaged coffee beans, can negatively affect aromas, which ultimately affects consumer experience and commercial trade [3,4].

Potato taste defect (PTD), which is unique to East African coffee, is characterized by a distinct musty, vegetable, raw potato aroma due to the presence of 2-isopropyl-3-methoxypyrazine (IPMP) in the volatile fraction of green coffee [5–7]. Field studies conducted in Burundi were the first to correlate the occurrence of PTD to the antestia bug (*Antestiopsis orbitalis*), which is native to East Africa [8]. The antestia bug feeds on various parts of the coffee plant like flower buds, berries, green shoots, and leaves, which results in an estimated yield loss of up to 40 % [6,9–11]. However, the mechanism linking insect infestation to high concentration of IPMP in coffee beans is still under investigation since the bug itself does not contain a

---

This chapter is reproduced from C. N. Cain, N. J. Haughn, H. J. Purcell, L. C. Marney, R. E. Synovec, C. T. Thoumsin, S. C. Jackels, K. J. Skogerboe, Analytical Determination of the Severity of Potato Taste Defect in Roasted East African Arabica Coffee, J. Agric. Food Chem. 7 (2021) 2253-2261.

measurable concentration of IPMP [6]. Microbial studies of coffee beans affected by PTD showed that bacteria belonging to two genera were responsible for formation of IPMP [12,13]. Therefore, it is hypothesized that the bug is either a vector for these microorganisms or damage from the antestia bug provides the optimal conditions for bacterial growth [12,13]. Another line of evidence suggests that antestia predation triggers an increase in *O*-methyltransferase (OMT) expression, which is responsible for converting precursor hydroxypyrazines into methoxypyrazines like IPMP and 2-isobutyl-3-methoxypyrazine (IBMP) [14]. Both mechanisms can explain the occurrence of PTD in coffee beans with and without visible insect damage [7].

Alterations to the sensory properties of coffee due to the presence of defects have been observed in a coffee sample's chemical profile, especially in the vapor headspace [4,15]. Headspace-solid phase microextraction gas chromatography-mass spectrometry (HS-SPME-GC-MS) is the prominent technique for studying the volatile analyte composition of defective coffee beans because of its sensitivity in extraction and detection [3,4,6,15]. Recently, HS-SPME-GC-MS was utilized to quantify the natural concentration of IPMP and IBMP in green coffee bean samples from different botanical species and geographical origin [7]. However, this study only analyzed two suspected PTD samples, where one batch was from Ethiopia and the other from Rwanda [7]. Surprisingly, the overall concentration of methoxypyrazines in these two suspected PTD batches were consistent with the natural baseline established for the geographic region and only inter-batch variability in the concentration of IPMP was seen for the suspected beans from Rwanda [7]. Given that these recent studies have focused on quantifying these methoxypyrazines in green coffee beans, the effect of PTD on roasted coffee beans remains an open question. Furthermore, due to the sensitivity and chemical selectivity provided by HS-SPME-GC-MS [16],

the total volatile analyte profile of roasted coffee beans affected by PTD can be closely examined to provide additional insight into this understudied defect.

The aim of this current study was to understand how PTD affects the chemical composition of roasted arabica coffee using both targeted and non-targeted approaches. Building upon and reinforcing previous work [7], this study assesses if the concentration of IBMP and IPMP can discriminate between coffee bean samples with and without PTD and ascertain if those concentrations are associated with the severity of odor attributed to PTD. To comprehensively make these determinations, high quality ground roasted coffee bean samples were sourced from three countries impacted by the occurrence of PTD: Burundi, Rwanda, and Uganda. Olfactory analysis classified the samples into four different categories: clean (i.e., no recognizable off-odor), mild PTD, medium PTD, and strong PTD. To provide optimal precision and accuracy of the HS-SPME-GC-MS analysis, these methoxypyrazines were quantified using a stable isotope dilution assay (SIDA) [17] in combination with the method of standard additions. Next, a supervised, non-targeted chemometric technique known as Fisher ratio (F-ratio) analysis [18–20] was applied to all the total ion chromatographic separations to discover additional analytes for which the concentrations differed based on the presence of PTD. Ultimately, the current study provides further knowledge of the role of methoxypyrazines and other volatile analytes in roasted East African arabica coffee beans affected by PTD.

2.2. Methods and Materials

2.2.1. Coffee samples and olfactory assessment

Specialty arabica coffee beans were sourced from Burundi, Rwanda, and Uganda by Counter Culture Coffee (Durham, NC, USA). Working under the assumption that PTD may be present to some degree in the samples [21], these coffee lots were found to be representative of

specialty coffee from this region and had no bulk PTD odor. The coffee lots were roasted to a light-medium degree, corresponding to a 11 – 13 % total weight loss and Agtron color scale score of 77 – 80 (Agtron Inc., Reno, NV, USA), on a gas powered roaster (Probat, Germany). Each individual coffee lot was then ground into ten-gram samples for a total of 3,800 ten-gram samples. These samples were subjected to olfactory assessment by trained staff at Counter Culture Coffee, using a method that has been described elsewhere [21]. Briefly, approximately 10 g of roasted coffee beans were ground, and the odor of the sample was classified into four categories: clean (i.e., no recognizable off-odor), mild PTD, medium PTD, and strong PTD. Samples suspected to have PTD were confirmed by a standard cupping protocol [22]. Of the 3,800 samples, 36 samples were found to have varying levels of PTD, illustrating how PTD odor becomes evident in smaller sample sizes. Of the remaining 3,764 samples having no detectable PTD odor, 13 representative "clean" samples were selected from the same lots for comparison. A total of 49 samples were then forwarded for analytical characterization.

2.2.2. Sample preparation

All chemical standards were obtained from Fisher Scientific or Sigma-Aldrich (USA). Sample preparation of roasted coffee beans is similar to previous studies, which quantified 2-alkyl-3-methoxypyrazines in juices and wines [23,24], with some minor modifications. Samples were composed of 1.5 g sodium chloride, 4.5 mL of ethylenediaminetetraacetic acid (EDTA; 0.111 M), 0.1 mL of internal standard deuterated-2-isobutyl-3-methoxypyrazine (d_3-IBMP; 150 ng/mL) and 0.1 g of ground roasted coffee in a 20 mL amber SPME vial. For standard additions samples, 0.2 mL IBMP stock (30 ng/mL) and 0.2 mL IPMP stock (30 ng/mL) in methanol (MeOH) were added. For non-spiked samples, 0.4 mL MeOH was added to equilibrate the matrices. EDTA was added to prevent thiol oxidation and reactions since volatile sulfur

compounds are pertinent to the aroma of coffee [25] and their concentrations may be influenced by the presence of PTD.

Initial screening demonstrated that 0.1 g of coffee was sufficient to ensure reproducible analysis over a broad range of IPMP concentrations without typically requiring further dilutions for samples with high levels of IPMP. The relative standard deviation (*RSD*) of IPMP concentration for replicate samples was 7.1 %. The slope of the IPMP standard addition data for 20 coffee samples was also used as a quality control measure to look for possible matrix interferences. Any sample where the slope fell outside of 0.51 ± 0.22 area/ng IPMP was flagged for repeat analysis. Additionally, the IPMP concentration of 60 % of the samples with an odor attribution were reanalyzed including all samples that appeared to be outliers. All of these were within 13 % of the original concentration determination, so the initial result was retained as the reported result. One strong sample was found to have a very high IPMP (529 ng/g) concentration and an outlier slope of 0.1. This sample required dilution in clean coffee to estimate the concentration of IPMP. The determination was performed at three different dilution ratios (1:2, 1:4 and 1:10) achieved by adding 0.1 g of this strong coffee sample to either 0.1, 0.3 or 0.9 g of previously screened clean ground coffee. Each of these dilutions was conducted in duplicate. The results were within 10 % of each other with an acceptable slope, resulting in confidence in the reported concentration.

The SPME fiber utilized was a divinylbenzene/carboxen/polydimethylsiloxane (DVB/CAR/PDMS) 50/30 μm (Supelco, USA). While previous work targeting methoxypyrazines in green coffee utilized a CAR/PDMS fiber [7], a DVB/CAR/PDMS sorbent was chosen here to maximize the molecular weight range and types of analytes extracted for F-ratio analysis while simultaneously focusing on the quantitation of IPMP and IBMP. Prior to first

use, the SPME fiber was conditioned in the injection port of the GC-MS at 250 ºC for a minimum of 30 min. At the start and end of each run reagent blanks were analyzed and showed no significant fiber contamination or carryover confirming effective conditioning and cleaning. The headspace of each prepared spiked and non-spiked sample was then extracted at 45 ºC for 60 min with constant agitation in an automated SPME injection system (CTC Analytics, CombiPAL). The agitator was set to 250 rpm with on/off times of 5/2 s. This extraction temperature and time was chosen based on previous work, which found these values to provide maximum chromatographic responses for pyrazines and other key volatile analytes [23].

2.2.3. Chromatographic conditions

The SPME fiber was then desorbed in the injection port of the GC-MS (Agilent 6890/5972, Agilent Technologies) at 250 ºC for 15 min. Separations occurred in splitless mode on a Restek Stabilwax column (60 m, 0.25 mm i.d., 0.5 μm film thickness) with a helium carrier gas flow rate of 1 mL/min. The GC oven was held at 40 ºC for 11 min, ramped to 190 ºC at 3 ºC/min, held at 190 ºC for 10 min, and then ramped to 250 ºC at 15 ºC/min for a total run time of 75 min. While the temperature program could have been optimized for the sole analysis of the two methoxypyrazines, this program was chosen to maximize the number of resolvable peaks for the F-ratio analysis involving the entire separations. Mass spectra were collected at 4.6 Hz between mass channels 50 – 350 *m/z* using electron impact ionization at 70 eV.

2.2.4. Data analysis

The chromatographic data was also imported into Matlab 2019b (Mathworks, Inc., Natick, MA, USA) by converting Agilent *.D files into Matlab *.m files with an in-house algorithm. The chromatographic data was unskewed to correct for retention time shifts caused by the scanning mass spectrometer [26] and baseline corrected using a rolling ball minimum

algorithm [27]. A localized retention time alignment algorithm was utilized to minimize the chromatogram-to-chromatogram variation observed [28]. First, the methoxypyrazines in the coffee samples were quantified using a targeted-based approach. IPMP, IBMP, and d_3-IBMP were confirmed in each chromatogram through retention time and mass spectra matching to a pure standard. Quantitation of the peak areas for IPMP, IBMP, and d_3-IBMP were carried out using their respective extracted ion current (EIC) chromatograms at *m/z* 137, 124, and 127. Calibration curves for IPMP were generated by normalizing the measured peak area of IPMP in the spiked samples to that of the internal standard. The presence of outliers in each odor attribution category was detected using a Grubbs' test and a one-way analysis of variance (ANOVA) determined statistical significance.

Next, F-ratio analysis was performed as a supervised, non-targeted method to discover analyte-specific differences between the clean and strong PTD samples. Due to the presence of two outlier samples in the strong PTD class (discussed later), F-ratio analysis was performed using all 13 clean samples and the remaining 13 strong PTD samples. The traditional F-ratio (^{T}F-ratio) calculation was calculated as the ratio of the between-class variance to the pooled within-class variance, referred to as the standard F-ratio calculation in previous work [29]. Additionally, two separate class-normalized F-ratio calculations were performed by dividing the between-class variance by the within-class variance of one group, or the other group [29,30]. Herein, this calculation is referred to as either the clean-normalized (^{C}F-ratio) or strong-normalized (^{S}F-ratio) F-ratio depending upon the group used in the denominator. F-ratio analysis was performed for every data point above a signal-to-noise (*S/N*) threshold of 10 with at least 6 samples above the *S/N* threshold. The average F-ratios were calculated using 3 *m/z* that produced the largest 3 F-

ratios. For F-ratios greater than the F-critical threshold, the analytes were arranged into "hit lists", ranking the analytes from the largest to smallest F-ratio.

For each analyte in each hit list, the peak area in each sample was quantified using the *m/z* associated with the top F-ratio, which has been shown to be selective, generally free from interferences [31]. A *t*-test ($\alpha = 0.05$) and concentration ratio between the clean and strong PTD classes ([Strong]/[Clean]) was also calculated. Tentative compound identifications were given if the mass spectrum was matched to a NIST11 library standard (National Institute of Standard and Technology, MD, USA) with a match value (MV) [32] greater than 800, which is standardly defined as sufficient to declare a match [33]. For unidentifiable analytes that were not statistically significant, the hit was labeled "Unk" and lettered in order of appearance in the hit lists, starting with the ^{T}F-ratio before the ^{C}F-ratio and ^{S}F-ratio. Similarly, an unidentifiable hit that was statistically significant was labeled "Unk" and numbered in order of appearance. A one-way ANOVA ($\alpha = 0.05$) also determined statistical significance for the normalized peak areas between all four odor classes.

2.3. Results and Discussion

2.3.1. Targeted analysis of methoxypyrazines

Methoxypyrazines like IPMP and IBMP, known for providing tainted aromas and flavors in coffee [23,24], have been identified in green [34] and roasted coffee [35]. Figure 2.1A shows the normalized total ion current (TIC) chromatograms for representative clean (red) and strong PTD (blue) samples. Application of a peak finder to these chromatograms indicated that over 100 peaks separated in the extracted coffee headspace, illustrating the resolving power of GC-MS. While the entire chromatogram will be analyzed later to discover additional chemical differences possibly related to PTD, current discussion will focus on the elution time window for

the methoxypyrazines of interest. Figure 2.1B shows a zoom-in of four overlaid normalized TIC chromatograms in the elution time window for IPMP (42.5 min), d_3-IBMP (46.6 min), and IBMP (46.7 min). The four representative TIC chromatograms relate to the four different categories for which the odor of the coffee grounds could be classified as: clean (red), mild (yellow), medium (green), and strong (blue). Note that the chromatograms shown in Figure 2.1 were normalized to the internal standard, d_3-IBMP, to a relative peak height of 1. As Figure 2.1 illustrates, the peaks for the natural and deuterated analogues of IBMP overlap in the TIC chromatogram. However, the benefit of using the deuterated internal standard, d_3-IBMP, is that it possesses selective *m/z* relative to IPMP or IBMP, which can used for quantitation.

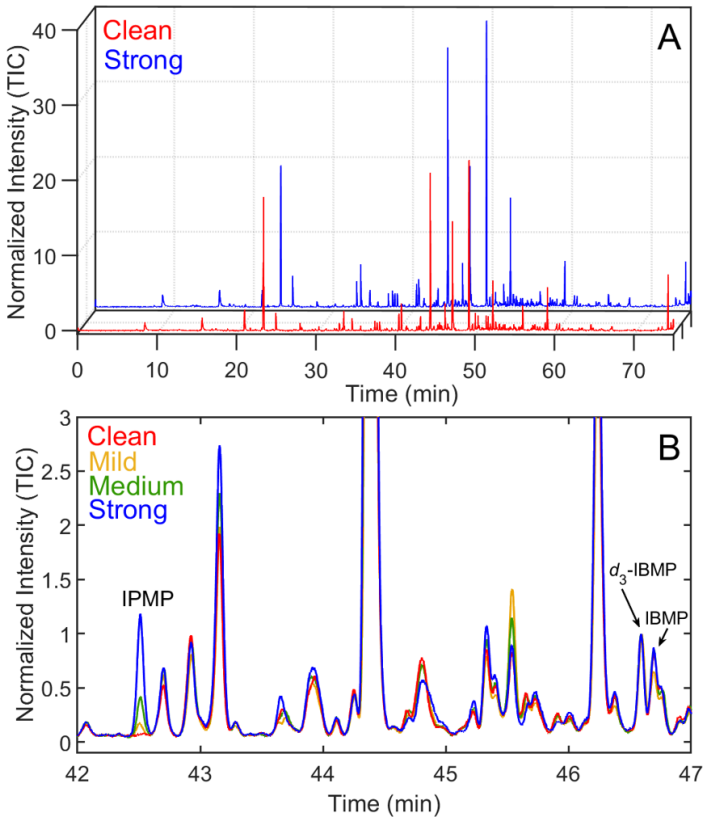


Figure 2.1. (A) Representative total ion current (TIC) chromatograms of a clean/non-PTD sample (red) and strong PTD (blue) sample. (B) Zoom-in of four normalized TIC chromatograms representing the different odor attributions of PTD to include the retention times associated with the methoxypyrazines of interest: clean (red), mild (yellow), medium (green), and strong (blue). Chromatographic peaks corresponding to IPMP, d_3-IBMP, and IBMP are labeled.

Figure 2.2 shows the normalized EIC chromatograms at *m/z* 124 (A) and 137 (B) for representative samples with the following odor attributions: clean (red), mild (yellow), medium (green), and strong (blue). Again, the chromatograms shown in Figure 2.2 were produced by normalizing the peak intensity of d_3-IBMP in the TIC chromatograms. When examining the mass spectra of IBMP and IPMP, *m/z* 124 and 137 are the base peak ions for each respective analyte. While IPMP can be seen at its retention time of 42.5 min in both EIC chromatograms, *m/z* 137 shows a greater detector response for IPMP than *m/z* 124. Therefore, these mass channels will not only provide the highest sensitivity to measure peak area, but they will also not experience any interferences from the internal standard. Visually, Figure 2.2A shows that, following normalization of the internal standard across all chromatograms, the IBMP peak area does not depend on the odor attributed to PTD; however, Figure 2.2B shows that the normalized IPMP peak area appears to vary based on the odor attributed to PTD. Comparing the variances calculated for the analytes of interest to the natural biological variation observed in food volatiles can further assess if there are relevant compositional differences in the coffee samples.

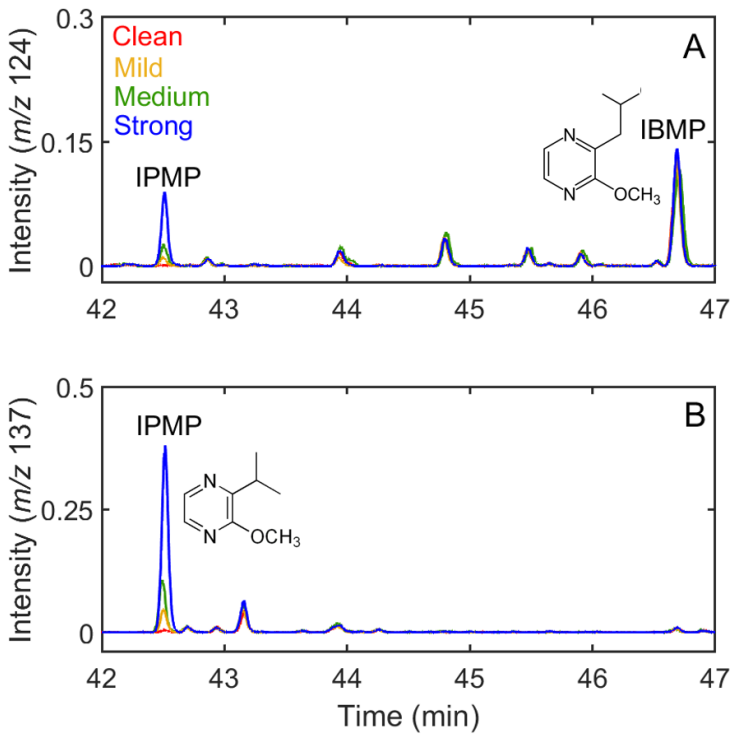


Figure 2.2. Normalized extracted ion current (EIC) chromatograms at *m/z* 124 (A) and 137 (B) for the measurement of IBMP and IPMP, respectively. Representative samples for clean (red), mild (yellow), medium (green), and strong (blue) PTD were chosen.

Previous studies indicated that the natural biological variation in volatile analyte concentrations for coffee and other food products can range from 20 – 50 % [36–40]. In the current study, the quantified peak areas for IBMP among all the samples have an *RSD* of 23.5 %, which is within the range of natural biological variations; thus, it seems that IBMP does not give insight into the presence of PTD. While IBMP is the primary odorant for green coffee beans and has measurable concentrations in roasted coffee beans, research suggests that its "vegetable-like" aroma is not perceived in roasted coffee beans, perhaps due to being masked by other characteristic coffee notes [35]. Furthermore, it was hypothesized that bug predation along with other environmental stresses may upregulate OMT expression in the coffee plant, contributing to elevated IBMP and IPMP levels in the coffee beans [14]. However, the concentrations of IBMP observed herein and previously [7] and lack of correlation to PTD suggests that other

environmental stressors may influence the expression of this OMT in all coffee plants. Even though IBMP may not aid in determining if roasted coffee beans have been affected by PTD, this analyte shows high variability based on geographical origin [7]. Therefore, future studies comparing methoxypyrazine concentration among different geographic origins, or more specifically countries, may want to consider the concentration relative to IBMP.

Meanwhile, the *RSD* in the normalized peak area of IPMP was calculated to be 197.7 %, which is significantly larger than the natural variation seen in food volatiles [36–40]. This result indicates that the variation in IPMP peak area is due to differences in the coffee beans induced by PTD instead of the biological variation between beans, further corroborating the results shown in Figure 2.2B. Therefore, the concentration of IPMP was quantified for all 49 African coffee samples to determine if it can be statistically linked to the odor attributed to PTD (i.e., clean, mild, medium, or strong). Quantitative analyses with HS-SPME rely upon a rigorous standardization method [41]. Use of an in-fiber standardization method, where the internal standard is pre-loaded onto the fiber coating, has been shown to provide remarkable reproducibility and successfully correct for matrix effects [42] and for the analysis of coffee volatiles [43,44]. While this method can be automated, quantitation of IPMP was performed herein using standard additions with SIDA which has also been shown to be reproducible and precise [17]. Calibration curves for the 49 samples were generated by measuring the ratio of IPMP (*m/z* 137) peak area to d_3-IBMP (*m/z* 124) peak area versus spiked IPMP concentration. Combining SIDA and the standard addition method resulted in a reproducibility of 7.1 % for the quantitation of IPMP, which is in agreement with previous work [23]. The limit of detection (*LOD*) and quantitation (*LOQ*) were determined to be 0.3 ng/g and 1.0 ng/g, respectively, indicating that the analytical methodology herein was reproducible and precise for analytes at

low concentrations [17,23]. Table A.1 provides the measured IPMP concentration and odor attribution for all 49 samples.

Prior to correlating IPMP concentration with PTD odor severity, Figure 2.3 shows a histogram comparing IPMP concentration in the clean samples (red) versus all the PTD affected samples group together (black). The inset of Figure 2.3 illustrates the wide range of IPMP concentrations (0.6 – 529.9 ng/g) tabulated for the 49 analyzed coffee samples. For the clean samples, the tabulated IPMP concentrations are all less than 4 ng/g with most of the samples having a concentration between 1 – 2 ng/g. This result is in agreement with previous work that showed the IPMP concentration would range between 1 – 3.2 ng/g regardless of the geographical origin of the green coffee beans [7]. Further reinforcing that there is a natural baseline level of IPMP in coffee samples, which could be due to the presence of other environmental stressors that upregulate the expression of OMTs and in turn IPMP [7,14]. Except for one sample with a concentration of 1.6 ng/g, the defective coffee samples had IPMP concentrations greater than 4 ng/g. While the concentration of IPMP is variable in the defective coffee samples, Figure 2.3 illustrates that samples affected by PTD could be distinguished based on IPMP concentration.

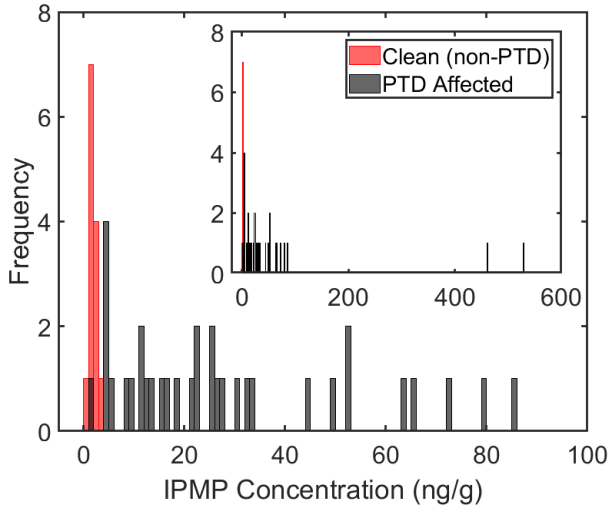


Figure 2.3. Distribution of IPMP concentrations in coffee beans that had either clean (red) or PTD (black) odor attributions. The inset figure depicts the wider range of IPMP concentrations.

Next, the IPMP concentration for each coffee sample was linked to its corresponding odor attributed to PTD. Using statistically excluded outliers, Table 2.1 reports the summary statistics for IPMP in the roasted coffee samples while Figure 2.4 provides a box-and-whiskers plot for the data. These results illustrate that both the median and average IPMP concentration increases as severity of odor attributed to PTD increases. The average concentration for IPMP in the clean, mild, medium, and strong PTD classes are 2.0 ± 0.2 ng/g, 9.5 ± 2.0 ng/g, 28.5 ± 6.2 ng/g, and 37.4 ± 6.7 ng/g, respectively. A one-way ANOVA test determined that there was a statistical difference in IPMP concentration among the four odor attributions ($p < 0.05$). Figure A.1 and Table A.2 provide the average IPMP concentration for each class with the addition of the outlier samples in the mild PTD odor class (72.4 ng/g) and strong PTD odor class (461.1 ng/g and 529.9 ng/g). Many studies investigating defects in coffee beans have focused on whether these defects can be discriminated against based on the presence or quantity of specific analytes, including work on PTD [1,3–7,15,45]. In recent work quantifying IPMP, it was briefly noted that the sensory characterization showed different degrees of defect intensity depending on the aliquot of one PTD sample [7]. Therefore, illustrating that IPMP concentration can not only predict the presence of PTD, but also the intensity of the odor attributed to the contamination of the coffee by defective beans.

Table 2.1. Summary statistics for IPMP concentration in each odor attribution category after outlier removal.

Odor Attribution	Number of Samples	Range (ng/g)	Median (ng/g)	Average (ng/g)	Standard Error (ng/g)
Clean	13	0.6 – 3.1	1.8	2.0	0.2
Mild	8	1.6 – 16.9	10.5	9.5	2.0
Medium	12	4.9 – 79.8	26.1	28.5	6.2
Strong	13	4.1 – 85.9	26.3	37.4	6.7

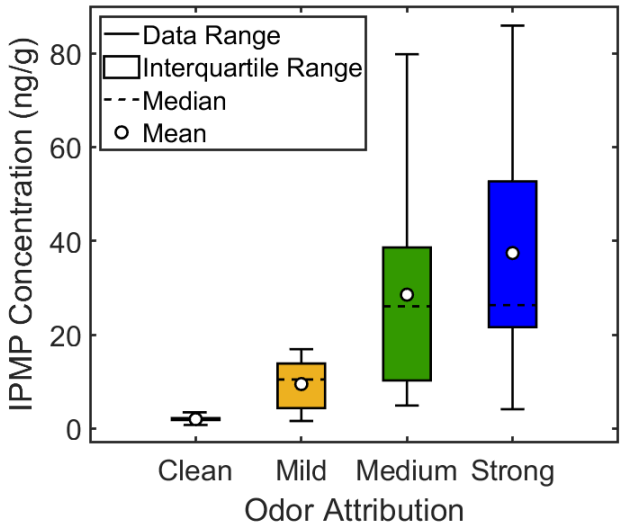


Figure 2.4. Box-and-whisker plots relating IPMP concentration to the intensity of odor attributed to PTD after statistically removing outliers.

2.3.2. Non-targeted analysis of PTD affected samples

Since all the volatiles present in the headspace can contribute to the final aroma [4,15,40,43,46], additional analytes were investigated to determine if their presence and intensity were related to the odor attributed to PTD in roasted East African coffee beans. As shown in Figure 2.1A, over 100 peaks were separated using the HS-SPME-GC-MS method herein. However, discovery of significant chemical differences can be hidden by insignificant chemical signals and background noise. Furthermore, it can be time-consuming to manually evaluate each peak in the chromatograms shown in Figure 2.1A. Therefore, F-ratio analysis was applied to discover differences in GC-MS data based on class membership with minimal analyst intervention [18–20].

F-ratio analysis was applied to the clean and strong PTD classes in a pixel-based approach on the aligned chromatograms, by calculating the F-ratio on a per-*m/z* basis at every data point pixel in the two-dimensional GC-MS matrix. The top 3 *m/z* with the highest F-ratio were averaged together at each data point pixel and peaks with F-ratios below the F-critical value

($\alpha = 0.05$) were set to zero. Figure 2.5A shows the traditional F-ratio (^{T}F-ratio) as a function of retention time. For the ^{T}F-ratio calculation, analytes that exhibit a high between-class variance relative to the pooled within-class variance are discovered as a hit. For example, IPMP (Figure 2.5B) is easily observed in the strong PTD samples (blue) and not seen in the clean samples (red), which indicates that the between-class variance is larger than the within-class variance. Figure 2.5C confirms this hypothesis since a sizeable ^{T}F-ratio of 22.9 can be seen at the retention time of IPMP. Since this peak has the largest ^{T}F-ratio seen for the comparison (Figure 2.5A), this analyte is assigned as the top hit in the ^{T}F-ratio hit list (Table A.3). The ^{T}F-ratio analysis discovered 41 peaks that could be potentially class-distinguishing, reducing the number of peaks for in-depth analysis by approximately half. Ultimately, applying this discovery-based technique can focus efforts in developing a volatile fingerprint for PTD.

F-ratio analysis can potentially be limited if a chemically meaningful analyte exhibits a large within-class variance or non-normal distribution [29]. Application of a class-normalized F-ratio calculation can discover these features that exhibit a high between-class variance relative to the within-class variance of one group [29,30]. This calculation can be potentially useful given the inherent biological variation present in foods like coffee due to growing conditions, genetics, and post-harvest processing [36–40]. The ^{C}F-ratio, which ratios the between-class variance to the within-class variance of the clean samples, discovered 49 analytes of interest (Table A.4). Similarly, the ^{S}F-ratio, which uses the within-class variance of the strong PTD samples, discovered 48 potentially class-distinguishing volatile analytes (Table A.5).

To consolidate these three hit lists (Table A.3 – Table A.5) into the volatile analyte fingerprint of PTD, each analyte peak was quantified using the *m/z* associated with the largest F-ratio since it is generally free of interferences [31]. The peak areas in the clean and strong PTD

odor classes were compared with a *t*-test ($\alpha = 0.05$) and analytes found to be class-distinguishing are listed in Table 2.2. Overall, 22 analytes were found to be class-distinguishing between the two odor classes and one of which was unable to be confidently identified with a MV above 800. Many of the volatile analytes in Table 2.2 were first discovered using the ^{T}F-ratio analysis (19 out of 22). However, three analytes were unique to the class-normalized F-ratio analyses: 4-vinyl guaiacol, 2-ethyl-3-methylpyrazine, and 2,6-diethylpyrazine. Discovery of these additional class-distinguishing analytes highlight the complementary nature of the ^{T}F-ratio and class-normalized F-ratio analyses and ensure completeness in developing a chemical profile of PTD.

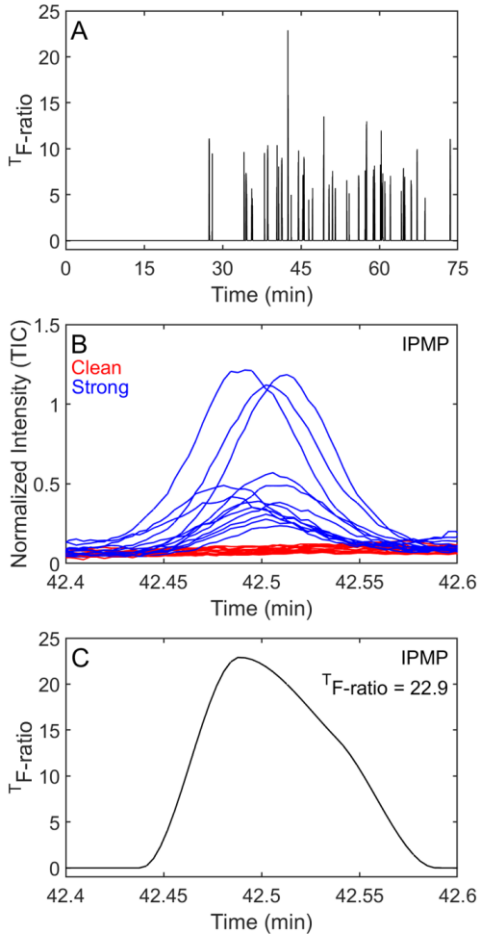


Figure 2.5. (A) The traditional F-ratio (^{T}F-ratio) calculated for the clean versus strong PTD comparison as a function of retention time. (B) Overlaid normalized TIC chromatograms of the IPMP peak for clean (red) and strong PTD (blue) samples. (C) Zoom-in of (A) to highlight the ^{T}F-ratio at the retention time of IPMP.

Table 2.2. Sensory description, F-ratio (hit number and value), and concentration ratio for analytes discovered to be statistically different ($p < 0.05$) between the clean and strong PTD samples.[a]

Compound	Sensory Description	F-ratio	[Strong]/[Clean]
Furfuryl formate	Ethereal, floral[2,37,48]	# 14 / 9.10	0.36
1-Methyl-1H-pyrrole	Smoky, woody, herbal[2,37,48]	# 5 / 11.1	0.63
Ethyl pyrazine	Nutty, roasted, cocoa[37,48]	# 12 / 9.54	0.71
2,3-Dimethylpyrazine	Nutty, caramel, cocoa[37,48]	# 8 / 10.4	0.72
3-Methylphenol	Phenolic[48]	# 13 / 9.47	0.74
2-Ethyl-3-methylpyrazine	Roasted, nutty[2,48]	# 41 / 5.79 [b]	0.74
3,5-Diethyl-2-methylpyrazine	Coffee, nutty[37,48]	# 25 / 7.13	0.77
2,6-Diethylpyrazine	Hazelnut, toasted[2,37,48]	# 37 / 6.64 [c]	0.80
2-Ethyl-5-methylpyrazine	Coffee, nutty, roasted[2,37,48]	# 7 / 10.4	0.80
1H-Pyrrole-2-carboxaldehyde	Musty, beefy, pungent[37,48]	# 9 / 9.93	1.42
1,6-Dihydrocarveol	Minty, camphoreous, terpenic[48]	# 29 / 6.56	1.65
3-Phenylfuran	Green bean-like[47]	# 16 / 8.26	1.71
1-Furfurylpyrrole	Plastic, vegetable, potato[48]	# 17 / 8.15	1.71
Unk1	N/A	# 21 / 7.66	1.74
Methyl salicylate	Minty, phenolic, camphoreous[48]	# 3 / 12.9	1.75
Ethyl 4-ethoxybenzoate	N/A	# 6 / 11.1	1.95
4-Ethyl-2-methoxyphenol	Woody, medicinal, phenolic[3,48]	# 40 / 4.43	1.96
Difurfuryl ether	Coffee, mushroom, earthy[48]	# 28 / 6.93	2.15
2-Methoxyphenol	Phenolic, smoky, medicinal, savory, meaty, woody[2,48]	# 4 / 12.0	2.16
2-Acetylpyrrole	Musty, bitter, grassy[2,48]	# 19 / 7.89	2.30
4-Vinyl guaiacol	Dry, woody, amber, sweet[2,48]	# 3 / 31.3 [b]	2.46
IPMP	Peasy, vegetable, potato-like[5,7,48]	# 1 / 22.9	21.4

[a] The sensory description was determined from the literature references provided. F-ratio values provided are for the ^{T}F-ratio unless otherwise noted. The volatile analytes in Table 2 were also ranked in ascending order of their concentration ratio.
[b] These analytes were first discovered with ^{C}F-ratio calculation and that value is provided.
[c] This analyte was first discovered with ^{S}F-ratio calculation and that value is provided.

The analytes in Table 2.2 were also ranked in ascending order of their concentration ratio, [Strong]/[Clean], which was calculated as the average normalized peak area in the strong PTD class divided by the clean class. The odor descriptions provided in Table 2.2 were from the literature [2,5,7,43,47,48]. Nine analytes (furfuryl formate to 2-ethyl-5-methylpyrazine) were observed to decrease in abundance for samples with strong PTD odors (i.e., having [Strong]/[Clean] between 0.36 and 0.80). Six of the nine analytes are alkylpyrazines, which are known to significantly contribute to coffee aroma [35]. Examination of the sensory description for these nine analytes shows that they all contribute favorable aromas (e.g., nutty, cocoa, roasted) to coffee. Conversely, thirteen analytes (1,6-dihydrocarveol to IPMP) were found in increased abundance for the strong PTD samples, which have a [Strong]/[Clean] ranging from 1.42 to 21.4. Naturally, IPMP had the largest [Strong]/[Clean], as it has been shown herein and in previous work [5–7] to be the characteristic marker of PTD. The other volatile analytes with increased concentrations in the strong PTD samples represent a wide variety of compound classes and sensory descriptions, which are mainly undesirable in coffee [49]. The main sensory descriptions for these analytes can be classified as vegetable-like, savory, and medicinal. These categorizations potentiate the hypothesis that the increased abundance of these volatile analytes while decreased abundance of compounds with more desirable aromas may amplify the odor of PTD.

For the analytes in Table 2.2, their peak areas were quantified in all four odor classes (Table A.6) and normalized to the internal standard. This quantitative information was utilized to create the heat map in Figure 2.6, where each analyte listed in Table 2.2 (using the same order) is represented by a row and each PTD odor class is a column. For ease of visualization, the measured peak areas for each analyte were normalized to their mean among the four classes.

Therefore, dark blue and maroon represent the PTD odor classes with the minimum and maximum peak area per each analyte, respectively. Figure 2.6 highlights that all these analytes track with the severity of odor attributed to PTD. For the first nine analytes listed (furfuryl formate to 2-ethyl-5-methylpyrazine), the measured peak areas decrease as the odor attributed to PTD becomes more severe. Conversely, the last thirteen analytes (1,6-dihydrocarveol to IPMP) all show that concentration increases as the odor attributed to PTD becomes more intense. Notably, a one-way ANOVA ($\alpha = 0.05$) found the peaks areas of the analytes listed to be statistically different among all four PTD odor classes investigated (Table A.6). Figure 2.6 suggests that these trends in roasted East African coffee are dependent on the occurrence of PTD rather than natural variability between samples, furthering our hypothesis that these correlated trends appear to contribute to the severity of odor attributed to PTD.

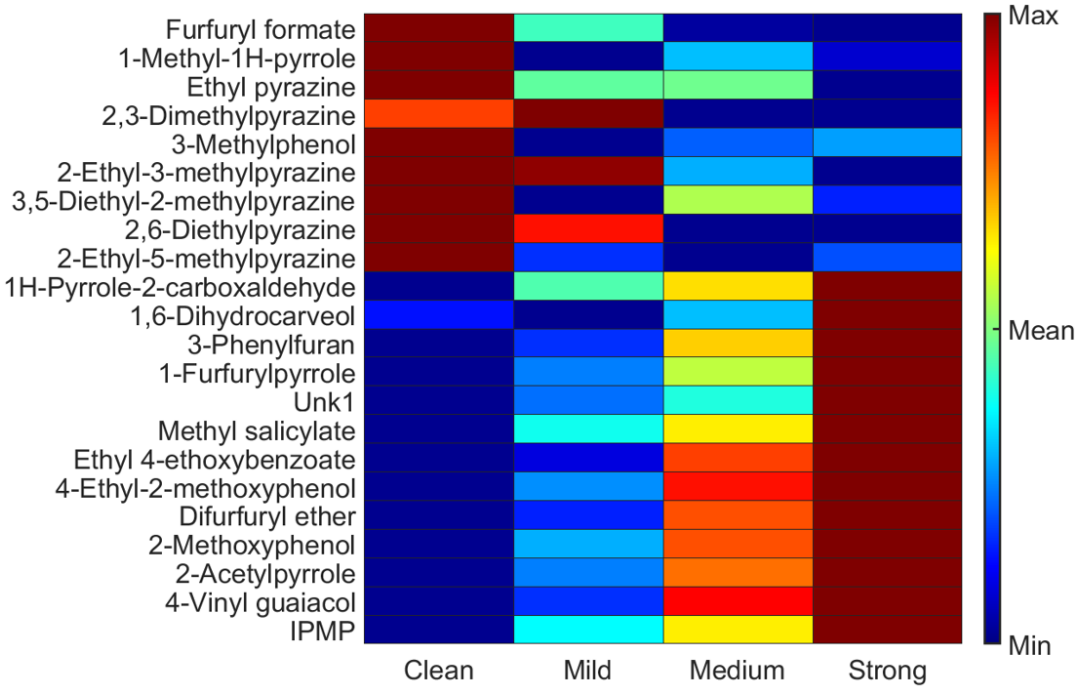


Figure 2.6. Heat map representing the normalized peak area measured for each analyte of interest (rows) in each PTD odor attribution class (columns). For each analyte, the odor class with the smallest peak area is shown as dark blue on the heat map and the odor class with the largest peak area is shown as maroon. The color bar (right) represents the scale used to represent the peak areas in each class for a given analyte.

Figure 2.7 shows the normalized EIC chromatograms for four analytes discovered by F-ratio analysis and listed in Table 2.2: furfuryl formate (A), ethyl pyrazine (B), 1-furfurylpyrrole (C), and 4-vinyl guaiacol (D). Chromatograms of both the clean (red) and strong PTD (blue) samples were plotted using the *m/z* with the largest F-ratio. These four analytes were chosen to represent the major chemical classes of volatiles in roasted coffee [2]. Furfuryl formate (Figure 2.7A) and ethyl pyrazine (Figure 2.7B) were found in decreased abundance in the strong PTD samples. However, 1-furfurylpyrrole (Figure 2.7C) and 4-vinyl guaiacol (Figure 2.7D) has increased abundance in the strong PTD samples. Previous research on Brazilian defective coffee beans, caused by poor agricultural practices and microbial growth, found decreased concentrations of furans and increased concentrations of pyrazines, pyrroles, and phenol derivatives [3,4,15]. These changes in the volatile profile of roasted Brazilian defective coffee beans are due to an internal imbalance of carbohydrates, amino acids, and chlorogenic acids during cultivation [15,50]. PTD has been associated with poor crop management practices [10,11] and microbial growth [13]. Therefore, internal alterations to the non-volatile fraction of green coffee beans, caused by PTD, can create the differences seen in the volatile profile of roasted coffee beans (Table 2.2; Figure 2.7). It is important to note that alkyl pyrazines were found in decreased abundance in the strong PTD samples (Figure 2.7B; Table 2.2) than predicted by the literature. This disagreement illustrates the difficulty in providing a full interpretation of the results herein since all these compound classes can be formed via different pathways during the roasting process. More research on how PTD alters the internal chemical composition of East African arabica coffee beans, which affects the volatile analyte profile after roasting, is necessary.

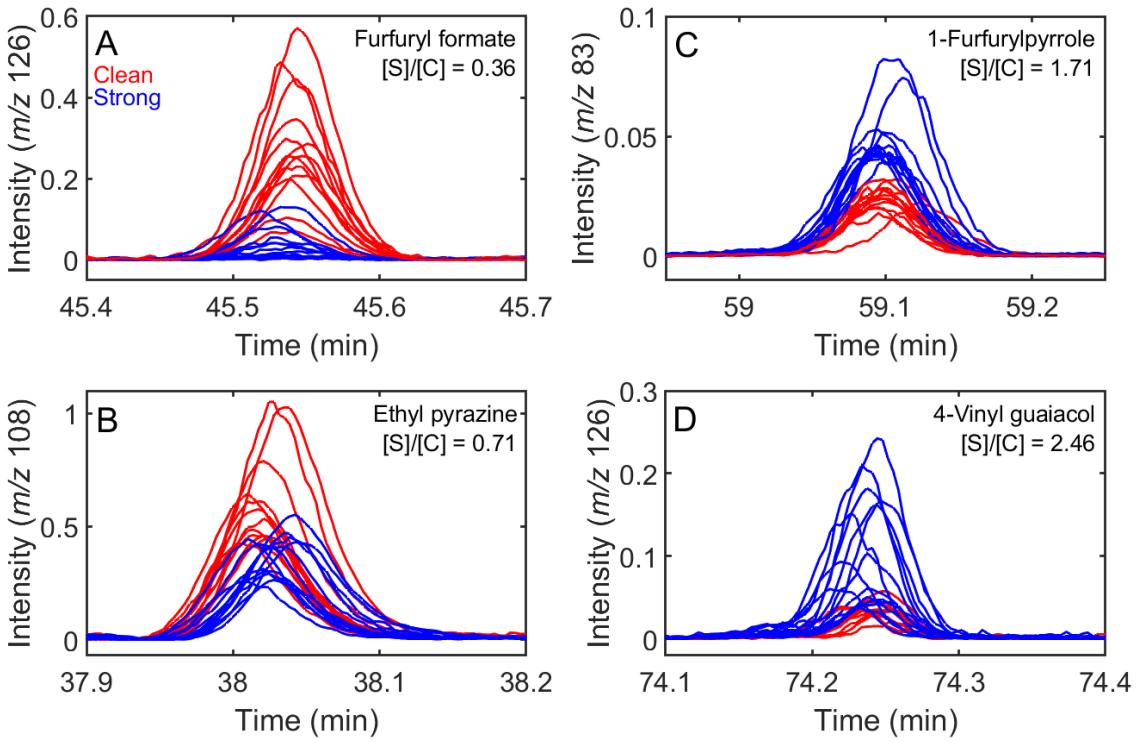


Figure 2.7. Overlaid normalized EIC chromatograms for analytes discovered by F-ratio analysis and were statistically different between the 13 clean (red) and 13 strong PTD (blue) samples. The concentration ratio between the strong and clean samples ([S]/[C]) is also provided. (A) Furfuryl formate at *m/z* 126. (B) Ethyl pyrazine at *m/z* 108. (C) 1-Furfurylpyrrole at *m/z* 83. (D) 4-Vinyl guaiacol at *m/z* 126.

2.4. Conclusion

In summary, the current study characterized the volatile analyte profile of roasted East African coffee beans affected by PTD using both targeted and non-targeted approaches. IBMP and IPMP were quantified using the former approach. Given the natural variation in IBMP among all 49 samples, regardless of the odor attributed to PTD, this investigation is in agreement with previous work [7] that this analyte alone cannot predict the presence of defective coffee beans. However, this report was the first to statistically show that the concentration of IPMP can be correlated to the severity of odor attributed to PTD and its detection via olfactory analysis works best for smaller segregated samples to limit dilution of the odor. Using a non-targeted chemometric method known as F-ratio analysis, 22 analytes (including IPMP) were discovered

to have statistically different concentrations between the clean and strong PTD samples. Volatiles with desirable aroma descriptions were found in decreased abundance in the strong PTD samples, whereas analytes with unpleasant aromas were found in increased abundance in the strong PTD samples. Comparison of each analyte among the four odor classes investigated statistically confirmed that these trends were due to the severity of PTD and not natural variability between samples. Furthermore, it is possible that PTD is caused by changes in the internal composition of non-volatiles in green coffee beans, which ultimately affects the volatiles produced during the roasting process. Future work is needed to further explore this hypothesis and better understand the mechanism of PTD in both raw and roasted coffee beans.

2.5. References

[1] N. Yang, C. Liu, X. Liu, T.K. Degn, M. Munchow, I. Fisk, Determination of volatile marker compounds of common coffee roast defects, Food Chem. 211 (2016) 206–214. https://doi.org/10.1016/j.foodchem.2016.04.124.

[2] A. Hameed, S.A. Hussain, M.U. Ijaz, S. Ullah, I. Pasha, H.A.R. Suleria, Farm to Consumer: Factors Affecting the Organoleptic Characteristics of Coffee. II: Postharvest Processing Factors, Compr. Rev. Food Sci. Food Saf. 17 (2018) 1184–1237. https://doi.org/10.1111/1541-4337.12365.

[3] P.D.C. Mancha Agresti, A.S. Franca, L.S. Oliveira, R. Augusti, Discrimination between defective and non-defective Brazilian coffee beans by their volatile profile, Food Chem. 106 (2008) 787–796. https://doi.org/10.1016/j.foodchem.2007.06.019.

[4] A.T. Toci, A. Farah, Volatile compounds as potential defective coffee beans' markers, Food Chem. 108 (2008) 1133–1141. https://doi.org/10.1016/j.foodchem.2007.11.064.

[5] R. Becker, B. Dohla, S. Nitz, O.G. Vitzthum, Identification of the "Peasy" Off-Flavour Note in Central African Coffees, in: 12th Int. Sci. Colloq. Coffee, Montreaux, Switzerland, 29 June - 3 July 1987, Association for Science and Information on Coffee (ASIC), Paris, France, 1987: pp. 203–215.

[6] S.C. Jackels, E.E. Marshall, A.G. Omaiye, R.L. Gianan, F.T. Lee, C.F. Jackels, GCMS investigation of volatile compounds in green coffee affected by potato taste defect and the antestia bug, J. Agric. Food Chem. 62 (2014) 10222–10229. https://doi.org/10.1021/jf5034416.

[7] D. Mutarutwa, L. Navarini, V. Lonzarich, P. Crisafulli, D. Compagnone, P. Pittia, Determination of 3-Alkyl-2-methoxypyrazines in Green Coffee: A Study to Unravel Their Role on Coffee Quality, J. Agric. Food Chem. 68 (2020) 4743–4751. https://doi.org/10.1021/acs.jafc.9b07476.

[8] B. Bouyjou, B. Decazy, G. Fourny, Removing the “potato taste” from Burundian Arabica, Plant. Rech. Dev. 6 (1999) 107–115.

[9] A.G. Ahmed, L.K. Murungi, R. Babin, Developmental biology and demographic parameters of antestia bug Antestiopsis thunbergii (Hemiptera: Pentatomidae), on Coffea arabica (Rubiaceae) at different constant temperatures, Int. J. Trop. Insect Sci. 36 (2016) 119–127. https://doi.org/10.1017/S1742758416000072.

[10] J. Bigirimana, A. Gerard, D. Mota-Sanchez, L.J. Gut, Options for Managing Antestiopsis thunbergii (Hemiptera: Pentatomidae) and the Relationship of Bug Density to the Occurrence of Potato Taste Defect in Coffee, Florida Entomol. 101 (2018) 580–586. https://doi.org/10.1653/024.101.0418.

[11] J. Bigirimana, C.G. Adams, C.M. Gatarayiha, J.C. Muhutu, L.J. Gut, Occurrence of potato taste defect in coffee and its relations with management practices in Rwanda, Agric. Ecosyst. Environ. 269 (2019) 82–87. https://doi.org/10.1016/j.agee.2018.09.022.

[12] D. Gueule, G. Fourny, E. Ageron, A. Le Flèche-Matéos, M. Vandenbogaert, P.A.D. Grimont, C. Cilas, Pantoea coffeiphila sp. nov., cause of the ‘potato taste’ of Arabica coffee from the African great lakes region, Int. J. Syst. Evol. Microbiol. 65 (2015) 23–29. https://doi.org/10.1099/ijs.0.063545-0.

[13] J.B. Ndayambaje, A. Nsabimana, S. Dushime, F. Ishimwe, H. Janvier, M.P. Ongol, Microbial identification of potato taste defect from coffee beans, Food Sci. Nutr. 7 (2019) 287–292. https://doi.org/10.1002/fsn3.887.

[14] K.E. Frato, Identification of Hydroxypyrazine O-Methyltransferase Genes in Coffea arabica: A Potential Source of Methoxypyrazines That Cause Potato Taste Defect, J. Agric. Food Chem. 67 (2019) 341–351. https://doi.org/10.1021/acs.jafc.8b04541.

[15] A.T. Toci, A. Farah, Volatile fingerprint of Brazilian defective coffee seeds: Corroboration of potential marker compounds and identification of new low quality indicators, Food Chem. 153 (2014) 298–314. https://doi.org/10.1016/j.foodchem.2013.12.040.

[16] Z. Zhang, J. Pawliszyn, Headspace Solid-Phase Microextraction, Anal. Chem. 65 (1993) 1843–1852. https://doi.org/10.1021/ac00062a008.

[17] P. Schieberle, W. Grosch, Quantitative Analysis of Aroma Compounds in Wheat and Rye Bread Crusts Using a Stable Isotope Dilution Assay, J. Agric. Food Chem. 35 (1987) 252–257. https://doi.org/10.1021/jf00074a021.

[18] K.M. Pierce, J.L. Hope, K.J. Johnson, B.W. Wright, R.E. Synovec, Classification of gasoline data obtained by gas chromatography using a piecewise alignment algorithm combined with feature selection and principal component analysis, J. Chromatogr. A. 1096 (2005) 101–110. https://doi.org/10.1016/j.chroma.2005.04.078.

[19] N.E. Watson, M.M. VanWingerden, K.M. Pierce, B.W. Wright, R.E. Synovec, Classification of high-speed gas chromatography-mass spectrometry data by principal component analysis coupled with piecewise alignment and feature selection, J. Chromatogr. A. 1129 (2006) 111–118. https://doi.org/10.1016/j.chroma.2006.06.087.

[20] C.E. Freye, P.R. Bowden, M.T. Greenfield, B.C. Tappan, Non-targeted discovery-based analysis for gas chromatography with mass spectrometry: A comparison of peak table, tile, and pixel-based Fisher ratio analysis, Talanta. 211 (2020) 120668. https://doi.org/10.1016/j.talanta.2019.120668.

[21] C. Thoumsin, Data Collection Methodology and Accurate Instance Rate Determination in Coffees With Potato Taste Defect (PTD), Count. Cult. Coffee. (2019) 1–9. https://counterculturecoffee.com/wp-content/uploads/2020/04/CCC-PTD-Paper-Final.pdf.

[22] Specialty Coffee Association of America, Cupping Specialty Coffee, (2015) 1–10. http://www.scaa.org/PDF/resources/ cupping-protocols.pdf (Accessed (accessed September 17, 2020).

[23] S. Boutou, P. Chatonnet, Rapid headspace solid-phase microextraction/gas chromatographic/mass spectrometric assay for the quantitative determination of some of the main odorants causing off-flavours in wine, J. Chromatogr. A. 1141 (2007) 1–9. https://doi.org/10.1016/j.chroma.2006.11.106.

[24] Y.S. Kotseridis, M. Spink, I.D. Brindle, A.J. Blake, M. Sears, X. Chen, G. Soleas, D. Inglis, G.J. Pickering, Quantitative analysis of 3-alkyl-2-methoxypyrazines in juice and wine using stable isotope labelled internal standard assay, J. Chromatogr. A. 1190 (2008) 294–301. https://doi.org/10.1016/j.chroma.2008.02.088.

[25] D.L. Capone, R. Ristic, K.H. Pardon, D.W. Jeffery, Simple quantitative determination of potent thiols at ultratrace levels in wine by derivatization and high-performance liquid chromatography-tandem mass spectrometry (HPLC-MS/MS) analysis, Anal. Chem. 87 (2015) 1226–1231. https://doi.org/10.1021/ac503883s.

[26] C.G. Fraga, Chemometric approach for the resolution and quantification of unresolved peaks in gas chromatography-selected-ion mass spectrometry data, J. Chromatogr. A. 1019 (2003) 31–42. https://doi.org/10.1016/S0021-9673(03)01329-3.

[27] H.G. Schmarr, J. Bernhardt, Profiling analysis of volatile compounds from fruits using comprehensive two-dimensional gas chromatography and image processing techniques, J. Chromatogr. A. 1217 (2010) 565–574. https://doi.org/10.1016/j.chroma.2009.11.063.

[28] K.J. Johnson, B.W. Wright, K.H. Jarman, R.E. Synovec, High-speed peak matching algorithm for retention time alignment of gas chromatographic data for chemometric analysis, J. Chromatogr. A. 996 (2003) 141–155. https://doi.org/10.1016/S0021-9673(03)00616-2.

[29] S.E. Prebihalo, G.S. Ochoa, K.L. Berrier, K.J. Skogerboe, K.L. Cameron, J.R. Trump, S.J. Svoboda, J.K. Wickiser, R.E. Synovec, Control-Normalized Fisher Ratio Analysis of Comprehensive Two-Dimensional Gas Chromatography Time-of-Flight Mass Spectrometry Data for Enhanced Biomarker Discovery in a Metabolomic Study of Orthopedic Knee-Ligament Injury, Anal. Chem. 92 (2020) 15526-15533. https://doi.org/10.1021/acs.analchem.0c03456.

[30] C. Brownie, D.D. Boos, J. Hughes-Oliver, Modifying the t and ANOVA F Tests When Treatment Is Expected to Increase Variability Relative to Controls, Biometrics. 46 (1990) 259–266.

[31] G.S. Ochoa, S.E. Prebihalo, B.C. Reaser, L.C. Marney, R.E. Synovec, Statistical inference of mass channel purity from Fisher ratio analysis using comprehensive two-dimensional gas chromatography with time of flight mass spectrometry data, J. Chromatogr. A. 1627 (2020) 461401. https://doi.org/10.1016/j.chroma.2020.461401.

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

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

[34] W. Holscher, H. Steinhart, Aroma Compounds in Green Coffee, Dev. Food Sci. 37 (1995) 785–803. https://doi.org/10.1016/S0167-4501(06)80196-2.

[35] M. Czerny, W. Grosch, Potent Odorants of Raw Arabica Coffee. Their Changes during Roasting, J. Agric. Food Chem. 48 (2000) 868–872. https://doi.org/10.1021/jf990609n.

[36] Y. Tikunov, A. Lommen, C.H.R. De Vos, H.A. Verhoeven, R.J. Bino, R.D. Hall, A.G. Bovy, A novel approach for nontargeted data analysis for metabolomics. Large-scale profiling of tomato fruit volatiles, Plant Physiol. 139 (2005) 1125–1137. https://doi.org/10.1104/pp.105.068130.

[37] B. Bertrand, R. Boulanger, S. Dussert, F. Ribeyre, L. Berthiot, F. Descroix, T. Joët, Climatic factors directly impact the volatile organic compound fingerprint in green Arabica coffee bean as well as coffee beverage quality, Food Chem. 135 (2012) 2575–2583. https://doi.org/10.1016/j.foodchem.2012.06.060.

[38] G. Weingart, B. Kluger, A. Forneck, R. Krska, R. Schuhmacher, Establishment and application of a metabolomics workflow for identification and profiling of volatiles from leaves of Vitis vinifera by HS-SPME-GC-MS, Phytochem. Anal. 23 (2012) 345–358. https://doi.org/10.1002/pca.1364.

[39] H.T.M. Tran, C.A.C. Vargas, L. Slade Lee, A. Furtado, H. Smyth, R. Henry, Variation in bean morphology and biochemical composition measured in different genetic groups of arabica coffee (Coffea arabica L.), Tree Genet. Genomes. 13 (2017). https://doi.org/10.1007/s11295-017-1138-8.

[40] N. Caporaso, M.B. Whitworth, C. Cui, I.D. Fisk, Variability of single bean coffee volatile compounds of Arabica and robusta roasted coffees analysed by SPME-GC-MS, Food Res. Int. 108 (2018) 628–640. https://doi.org/10.1016/j.foodres.2018.03.077.

[41] J.A. Field, G. Nickerson, D.D. James, C. Heider, Determination of essential oils in hops by headspace solid-phase microextraction, J. Agric. Food Chem. 44 (1996) 1768–1772. https://doi.org/10.1021/jf950663d.

[42] Y. Wang, J. O'Reilly, Y. Chen, J. Pawliszyn, Equilibrium in-fibre standardisation technique for solid-phase microextraction, J. Chromatogr. A. 1072 (2005) 13–17. https://doi.org/10.1016/j.chroma.2004.12.084.

[43] D. Bressanello, E. Liberto, C. Cordero, B. Sgorbini, P. Rubiolo, G. Pellegrino, M.R. Ruosi, C. Bicchi, Chemometric Modeling of Coffee Sensory Notes through Their Chemical Signatures: Potential and Limits in Defining an Analytical Tool for Quality

Control, J. Agric. Food Chem. 66 (2018) 7096–7109. https://doi.org/10.1021/acs.jafc.8b01340.

[44] D. Bressanello, E. Liberto, C. Cordero, P. Rubiolo, G. Pellegrino, M.R. Ruosi, C. Bicchi, Coffee aroma: Chemometric comparison of the chemical information provided by three different samplings combined with GC–MS to describe the sensory properties in cup, Food Chem. 214 (2017) 218–226. https://doi.org/10.1016/j.foodchem.2016.07.088.

[45] A.S. Franca, L.S. Oliveira, J.C.F. Mendonça, X.A. Silva, Physical and chemical attributes of defective crude and roasted coffee beans, Food Chem. 90 (2005) 89–94. https://doi.org/10.1016/j.foodchem.2004.03.028.

[46] C. Scheidig, M. Czerny, P. Schieberle, Changes in key odorants of raw coffee beans during storage under defined conditions, J. Agric. Food Chem. 55 (2007) 5768–5775. https://doi.org/10.1021/jf070488o.

[47] Y. Feng, G. Su, H. Zhao, Y. Cai, C. Cui, D. Sun-Waterhouse, M. Zhao, Characterisation of aroma profiles of commercial soy sauce by odour activity value and omission test, Food Chem. 167 (2015) 220–228. https://doi.org/10.1016/j.foodchem.2014.06.057.

[48] The Good Scents Company, The Good Scents Company Information System, (2018). http://www.thegoodscentscompany.com/.

[49] I. López-Galilea, N. Fournier, C. Cid, E. Guichard, Changes in headspace volatile concentrations of coffee brews caused by the roasting process and the brewing procedure, J. Agric. Food Chem. 54 (2006) 8560–8566. https://doi.org/10.1021/jf061178t.

[50] A. Farah, M.C. Monteiro, V. Calado, A.S. Franca, L.C. Trugo, Correlation between cup quality and chemical attributes of Brazilian coffee, Food Chem. 98 (2006) 373–380. https://doi.org/10.1016/j.foodchem.2005.07.032.

Chapter 3: Investigating Sensory-Classified Roasted Arabica Coffee with GC×GC-TOFMS and Chemometrics to Understand Potato Taste Defect

3.1. Introduction

Since coffee quality is directly linked to taste and aroma, the presence of product defects due to agricultural practices and/or an adverse agricultural environment can cause off-flavors and devalue the crop [1–3]. Potato taste defect (PTD) is a sporadic defect that occurs in coffee beans grown in the African Great Lakes region, namely those cultivated in Burundi, Rwanda, and Uganda [4,5]. As the name implies, this defect is characterized by the distinct musty, potato aroma of affected coffee beans. The occurrence of this defect has been linked to the presence of *Antestiopsis orbitalis*, an antestia bug native to this region that feeds on the coffee plant [5–7]. Chemically, studies of both green and roasted coffee beans have discovered that the presence and strength of PTD is correlated to 2-isopropyl-3-methoxypyrazine (IPMP) [4,8–11]. However, the mechanism linking coffee plant damage from the antestia bug to the presence of IPMP in volatile headspace of affected beans remains unclear, especially since PTD can be present in coffee beans with and without visual insect damage [9]. It has been hypothesized that damage from the bug either provides favorable growth conditions for microorganisms that can produce IPMP [12–14] or initiates the conversion of hydroxypyrazines naturally produced by the plant into methoxypyrazines like IPMP via O-methyltransferase expression [15].

While IPMP has been chemically linked to PTD [4,8–11], it is important to note that the chemical composition of coffee is highly complex with numerous other volatiles contributing to

---

This chapter is reproduced from C. N. Cain, M. Gaida, P.-H. Stefanuto, J.-F. Focant, R. E. Synovec, S. C. Jackels, K. J. Skogerboe, Investigating Sensory-Classified Roasted Arabica Coffee with GC×GC-TOFMS and Chemometrics to Understand Potato Taste Defect, Microchem. J. 196 (2024), 109578.

aroma [16]. The presence of these other analytes could, in turn, either heighten or mask the odor of PTD. Volatile fingerprinting of coffee, especially of beans affected by PTD, has primarily been performed using one-dimensional (1D) gas chromatography-quadrupole mass spectrometry (GC-MS) [1–4,8–11]. For example, a recent study demonstrated that two key volatiles in roasted coffee aroma, 2-ethyl-3,5-dimethylpyrazine and 2-furfurylthiol, can mask the odor attributed to IPMP [11]. Other previous work applying headspace-solid-phase-microextraction GC-MS (HS-SPME-GC-MS) identified 22 analytes, including IPMP, whose signals differentiated samples that did not have a detectable off-odor ("clean" samples as scored by a sensory panel) from those affected by PTD [10]. Compounds with a higher abundance in the clean coffee samples generally had desirable aromas, whereas compounds linked to undesirable aromas had larger signals in samples affected by PTD [10]. The results of these studies highlight the potential for PTD to affect the concentration of other volatiles present in coffee.

However, 1D separation methods are inherently limited in their peak capacity, which constrains the number of analytes that can be resolved in a reasonable analysis time [17]. As a result, the discovery of volatiles impacted by the presence of PTD will also be limited due to the use of GC-MS. Fortunately, use of comprehensive two-dimensional gas chromatography with time-of-flight mass spectrometry (GC×GC-TOFMS) can readily increase the peak capacity of a 1D-GC separation. A GC×GC separation connects two complementary columns via a modulator, which continuously collects fractions of effluent from the first dimension (^{1}D) separation and reinjects those fractions on the second dimension (^{2}D) [18]. Due to this modulation process, the peak capacity of an ideal GC×GC separation is approximately 10-fold higher than its 1D-GC counterpart [19] and the sensitivity of the GC×GC separation is also enhanced [20]. The

advantages of this separation platform have been illustrated in various food analysis studies [21,22], including those aimed at profiling coffee aroma [23–28].

While the increased resolving power provided by GC×GC-TOFMS is beneficial for chemical fingerprinting studies, manual identification, and signal integration of every peak present in a chromatogram can be burdensome due to the size and complexity of the data. Fortunately, the data set produced by GC×GC-TOFMS analysis enables the use of non-targeted chemometrics, which can identify statistically significant chemical signals and elucidate relationships among different samples in an automated fashion. For example, GC×GC-TOFMS and non-targeted chemometrics have been coupled to profile commercial espresso capsules [26] and differentiate decaffeinated coffee from its regular coffee counterpart [27]. Fisher ratio (F-ratio) analysis has been one of the prominent methods utilized in food analysis studies for the discovery of class-distinguishing analytes [27,29–33]. Using a priori knowledge of the sample classes, the ratio of the between-class variance to the pooled within-class variance (a F-ratio) can be calculated for every data point [34], peak listed in a table [35], or a tile (binned data) in a chromatographic data set [36,37]. The output is a list of retention times for peaks ranked in descending order of their F-ratio, referred to as a "hit list". A peak with a large F-ratio is likely class-distinguishing since there is a large variance between classes relative to the within-class variance.

Implementation of F-ratio analysis should ensure that the discovery of chemically relevant differences is not hindered by retention time misalignment and spurious detector noise. The use of tiling (i.e., "smart binning") can prevent the discovery of these instrumental artifacts by dividing the chromatogram into small, rectangular tile sections which capture the entire peak signal along with any retention time shifting. Furthermore, tile-based non-targeted methods can

also improve the discovery of low-level peaks due to enhancements in the relative signal-to-noise (*S/N*) while also automatically removing redundant hits [36,37]. Given these advantages, a tile-based algorithmic platform has been recently commercialized and adapted for various experimental designs [38–41]. Herein, the capabilities of tile-based non-targeted F-ratio analysis are illustrated on a HS-SPME-GC×GC-TOFMS data set of PTD in roasted arabica coffee. Olfactory analysis categorized these samples into four classes based on odor severity: clean, mild PTD, medium PTD, and strong PTD [10]. Class-distinguishing analytes discovered by tile-based F-ratio analysis were quantified using a pure mass channel (*m/z*) identified for the analyte. With GC×GC, a nearly a 10-fold larger peak capacity is anticipated [19] and enhanced *S/N* due to the use of thermal modulation [20]; thus, giving rise to the hypothesis that a significantly larger number of compounds will be discovered relative to our previous 1D-GC-MS study [10]. The consequences of this improved discovery of relevant sensory compounds will be further discussed. Principal components analysis (PCA) and partial least squares (PLS) regression are also performed to illustrate how the analytes discovered by tile-based F-ratio analysis contribute to the biochemical understanding PTD in coffee beans from the African Great Lakes region.

3.2. Methods and Materials

3.2.1. Acquisition and assessment of coffee samples

Arabica coffee samples from Burundi, Rwanda, and Uganda were sourced, roasted, and initially assessed by Counter Culture Coffee (Durham, NC, USA). The green coffee beans were sourced according to standard procedures, ensuring similarities in their screen size and moisture content and lack of primary and secondary defects [42]. Details of the roasting conditions and olfactory evaluation of the ground coffee samples were previously reported [10,42]. Briefly, the green coffee beans were roasted to a light-medium roast, resulting in a weight loss of 11-13 %

and score of 77-80 on the Agtron color scale, and then ground. These coffee samples (56 total) were then classified by a sensory panel into one of four groups (clean, mild PTD, medium PTD, and strong PTD) based on a previously established protocol [42]. This protocol involved grinding 10 g portions of whole roasted coffee beans into a clean cup, carefully smelling the grounds, and categorizing them based on their odor intensity. Suspected PTD samples were then confirmed via cupping protocols published by the Specialty Coffee Association [43]. After olfactory analysis, a total of 14 clean samples, 11 mild PTD samples, 13 medium PTD samples, and 18 strong PTD samples were forwarded for analytical characterization (Table B.1; Figure B.1). The concentration of IPMP ranged from 0 ng/g – 3.1 ng/g for the clean samples, 1.6 ng/g – 72.4 ng/g for the mild PTD samples, 4.9 ng/g – 79.8 ng/g for the medium PTD samples, and 4.1 ng/g – 529.9 ng/g for the strong PTD samples. All samples were stored in airtight glass containers prior to sample preparation, extraction, and GC×GC-TOFMS analysis.

3.2.2. Sample preparation and extraction

All analytical standards were procured from Fisher Scientific or Sigma-Aldrich (USA). Preparation of the coffee samples for this study was similar to our previous methodology [10]. Ground roasted coffee (0.1 g) were added to a 20 mL SPME vial along with 1.5 g of sodium chloride, 4.5 mL of ethylenediaminetetraacetic acid (EDTA; 0.111 M), 480 µL of methanol, and 20 µL of the internal standard, deuterated-2-isobutyl-3-methoxypyrazine (d_3-IBMP; 5 ppm). A divinylbenzene/carboxen/polydimethylsiloxane (DVB/CAR/PDMS; fiber thickness: 50/30 µm; Restek, Bellefonte, PA, USA) SPME fiber was utilized for the extraction of the headspace volatiles. This SPME fiber was selected due to its popularity and effectiveness in coffee studies [2,3,8,10,23,26,27,44], where it has been shown to extract a wide range of molecular weights and compound classes. Initial conditioning of a new DVB/CAR/PDMS fiber followed the

manufacturer's guidelines, where it was held in a 270 ºC GC inlet for 1 hr. Prior to each sample extraction, the coffee samples were incubated at 60 ºC for 15 min while being agitated at 250 rpm with on/off times of 5 s and 2 s, respectively. The headspace of the coffee samples was extracted for 30 min at 60 ºC. Slightly different from previous work [10], this extraction temperature and time was found to provide an optimal balance between amplifying the chromatographic signal of IPMP while detecting other coffee volatiles. The SPME fiber with the extracted volatiles was desorbed splitless in the GC inlet for 5 min at 250 ºC. Between sample extractions and chromatographic runs, the fiber was re-conditioned at 270 ºC for 10 min. Note, chromatographic blanks were periodically collected to ensure the SPME fiber was properly cleaned in between sample extractions. These chromatographic blanks did not show the presence of peaks and/or contaminants that could affect the results of this study. The sample extraction method utilized the L-PAL3 autosampler (LECO, St. Joseph, MI, USA).

3.2.3. Chromatographic conditions

Separations of each coffee sample were collected in duplicate using the LECO Pegasus BT 4D GC×GC-TOFMS equipped with an Agilent 7890 GC (Agilent Technologies, Palo Alto, CA, USA) and a stock quad-jet thermal modulator. To prevent contamination and carryover from other experiments performed in the laboratory, a new set of conditioned GC columns was installed in the GC×GC-TOFMS, along with a new inlet liner. Splitless sample injections were separated on a polar Rtx-Wax ^{1}D column (30 m × 0.25 mm × 0.25 μm; Restek), and a non-polar Rxi-1MS ^{2}D column (1.7 m × 0.18 mm × 0.18 μm; Restek). An Rtx-Wax column was used in the ^{1}D to maintain a similar primary separation as in our previous 1D-GC-MS study, which also used this column type [10]. Furthermore, previous GC×GC studies on coffee volatiles have also used this GC×GC column configuration [23,27]. The ^{1}D column was held at 40 ºC for 5.5 min before

ramping to 240 ºC at 5 ºC/min, where it was held for 5 min. The same temperature program was used for the ^{2}D oven and modulator with an offset of +5 ºC and +15 ºC, respectively. The carrier gas, ultra-high purity helium (Grade 5, 99.999 %, Praxair, Seattle, WA, USA), operated at a constant flow rate of 1 mL/min. The ^{1}D effluent was reinjected on the ^{2}D column at a modulation period of 2 s. The ion source and transfer line temperatures were set to 225 ºC and 285 ºC, respectively. The TOFMS collected *m/z* 45-350 at 100 Hz with an electron ionization energy of 70 eV after a 10 s acquisition delay.

3.2.4. Data analysis

Following data acquisition, the chromatograms were imported into Matlab 2019b (Mathworks, Inc., Natick, MA, USA) for further analysis. The chromatograms were baseline corrected and normalized to the peak area of the internal standard, d_3-IBMP, at *m/z* 127. F-ratio analysis was performed by comparing the clean and strong PTD coffee samples. Note, since the number of samples in the clean and strong PTD class was unbalanced, 14 strong PTD coffee samples were arbitrarily selected for the F-ratio comparison. Hence, with 14 coffee samples per-class and two replicates were collected per-sample, a total of 28 chromatograms per-class were compared. A tile size of 10 s × 200 ms (^{1}D × ^{2}D) and cluster window size of 6 s × 120 ms was selected based on the peak widths and degree of retention time shifting observed in the chromatograms. The verification that these tile bin parameters were appropriate for these samples is provided in the next section. Using the four-grid schemes, F-ratios were calculated on every tile per-*m/z* that had a *S/N* greater than 10. The resulting hits were then ranked according to their top F-ratio *m/z*. Any remaining redundant hits and artifacts from the SPME fiber or column bleed were removed from the final hit list. As a result of this methodology, the final hit list

contained both class-distinguishing (i.e., true positives) and non-class-distinguishing (i.e., false positives) analytes.

To identify the class-distinguishing analytes in the hit list (i.e., the true positives), a *t*-test (assuming unequal variances) was calculated for every analyte discovered. A *p*-value was calculated using the signal encapsulated in the tile surrounding each hit at the top F-ratio *m/z*. Hits with a *p*-value < 0.01 from a *t*-test were considered as true positives, which were subjected to later identification and quantitation efforts, while hits with a *p*-value > 0.01 were false positives [33]. Next, for those analytes determined to be true positives, tentative compound identifications were determined by matching the acquired mass spectrum to the NIST 11 library (National Institute of Standards and Technology, Gaithersburg, MD, USA). A match value (MV) ≥ 800 was required for (tentative) identification [45]. Multivariate curve resolution-alternating least squares (MCR-ALS) was used to improve identification efforts for 283 analytes by resolving the hit mass spectrum from the background noise and interferences [46]. Hits that still could not be identified with a MV ≥ 800 after applying MCR-ALS are labeled as "unknown" and numbered. Lastly, for those true positive analytes, accurate concentration ratios between the clean and strong PTD samples, referred to as [Strong]/[Clean] herein, were determined using a pure *m/z* for each hit. Pure analyte *m/z* were discovered using a recent extension to the tile-based software known as the signal ratio (S-ratio) algorithm [47]. Two signal consistency metrics are implemented to discover pure *m/z* for the target analyte using two metrics, a lack-of-fit (*LOF*) and *p*-value from a *t*-test (assuming unequal variances). Ideally, a pure analyte *m/z* should have a significant difference in signals between two classes (i.e., a small *p*-value), but the peak shapes between classes should be similar (i.e., no significant *LOF* is identified). Therefore, the S-ratio calculated for a sufficiently pure analyte *m/z* accurately approximates the true concentration ratio

[47]. Meanwhile, *m/z* from an unresolved analyte(s) with little chemical variation between classes will have a high *p*-value, low *LOF*, and S-ratio equal to 1. Using these metrics, the purest *m/z* for quantitation had the lowest *LOF* (maximum *LOF* tolerated = 20 %) and *p*-value (maximum *p*-value tolerated = 1×10^{-4}). Every class-distinguishing analyte had at least one *m/z* that met the *LOF* and *p*-value thresholds to be labeled as sufficiently pure for quantitation. For improved visualization of solely the analytes discovered by F-ratio analysis, "stitch" GC×GC chromatograms were constructed. The principles of this visualization technique have been described previously [48,49]. Briefly, this method extracts the chromatographic signal within a 10 s × 200 ms tile at a pure analyte *m/z* for every class-distinguishing analyte. A *S/N* filter of 10 was implemented to remove noise from the extracted tiles. The tiles were then inserted back into an empty "chromatogram" at their original retention time locations.

Both PCA and PLS models were developed using these "stitch" chromatograms with PLS Toolbox 8.9 (Eigenvector Research, Manson, WA, USA). The signals from each sample, quantified at the pure analyte *m/*z, were mean-centered and inputted into PCA to demonstrate the differences between PTD odor classes after F-ratio analysis. The hits discovered by F-ratio analysis were also utilized to build a predictive model of IPMP concentration via PLS regression. For PLS modeling, the data was divided into a calibration and validation data set using the Kennard-Stone algorithm. A calibration model using 42 coffee samples was built using the mean-centered signal data and auto-scaled values for IPMP concentration. Venetian-blinds cross-validation with 6-splits was performed to determine the appropriate number of latent variables (LVs) for the model and calculate a normalized root-mean-square error of cross-validation (NRMSECV). The validation data set was then input into the PLS model to determine the NRMSE of prediction (NRMSEP).

3.3. Results and Discussion

The TIC chromatograms of a clean (A) and strong PTD (B) sample are shown in Figure 3.1, illustrating the complexity of the volatile headspace and need for a GC×GC separation to discover analytes indicative of PTD. A peak detection algorithm identified approximately 500 peaks present in these TIC GC×GC chromatograms, excluding artifacts such as streaks. The total peak capacity for these GC×GC separations was 3750 based on an average ^{1}D and ^{2}D width-at-base (W_b) of 8 s and 200 ms, respectively. For reference, the 1D-GC TIC chromatograms for these coffee samples had a peak capacity of 390, based on an average W_b of 10 s, and only identified ~100 peaks present [10]. The ~5-fold increase in the number of peaks resolved can be attributed to the increased peak capacity and *S/N* provided by a GC×GC separation. Visual comparison of the chromatograms shown in Figure 3.1 demonstrates that the strong PTD sample (B) has an overall reduction in signal compared to the intensity of the peaks present in the clean sample (A). The scale-expanded chromatograms between 16-22 min on the ^{1}D and 1-1.8 s on the ^{2}D (Figure 3.1C-D) further highlight the differences in these samples. Figure 3.1C-D denotes the locations of three representative analytes known to contribute to the aroma of coffee: (1) 2-ethyl-6-methylpyrazine, (2) 2,3-diethylpyrazine, and (3) linalool. Note, the peaks corresponding to IPMP and the internal standard (IS), d_3-IBMP, are also labeled in Figure 3.1C-D. Close examination of this region highlights not only the complexity of these chromatograms, but also noticeable concentration differences between the clean and strong PTD coffee samples. For example, alkylpyrazines like 2-ethyl-6-methylpyrazine (labeled as analyte 1) and 2,3-diethylpyrazine (labeled as analyte 2) are known to contribute a roasted, nutty aroma in the coffee headspace [44,50]. Linalool (labeled as analyte 3) has also been identified as one of the few alcohols pertinent to the odor of arabica coffee [3], providing a floral and fruity aroma in

roasted coffee [50]. However, Figure 3.1C-D shows that the signal for these analytes is ~1.5-fold to ~50-fold lower in the strong PTD sample compared to the clean coffee sample. Meanwhile, Figure 3.1C-D demonstrates that the signal for IPMP (outlined by the pink oval) becomes present in the strong PTD coffee sample, which is consistent with previous studies [9,10]. The chromatographic complexity highlighted in Figure 3.1 illustrates the utility in applying non-targeted chemometric methods to develop a "comprehensive" volatile fingerprint of PTD in coffee.

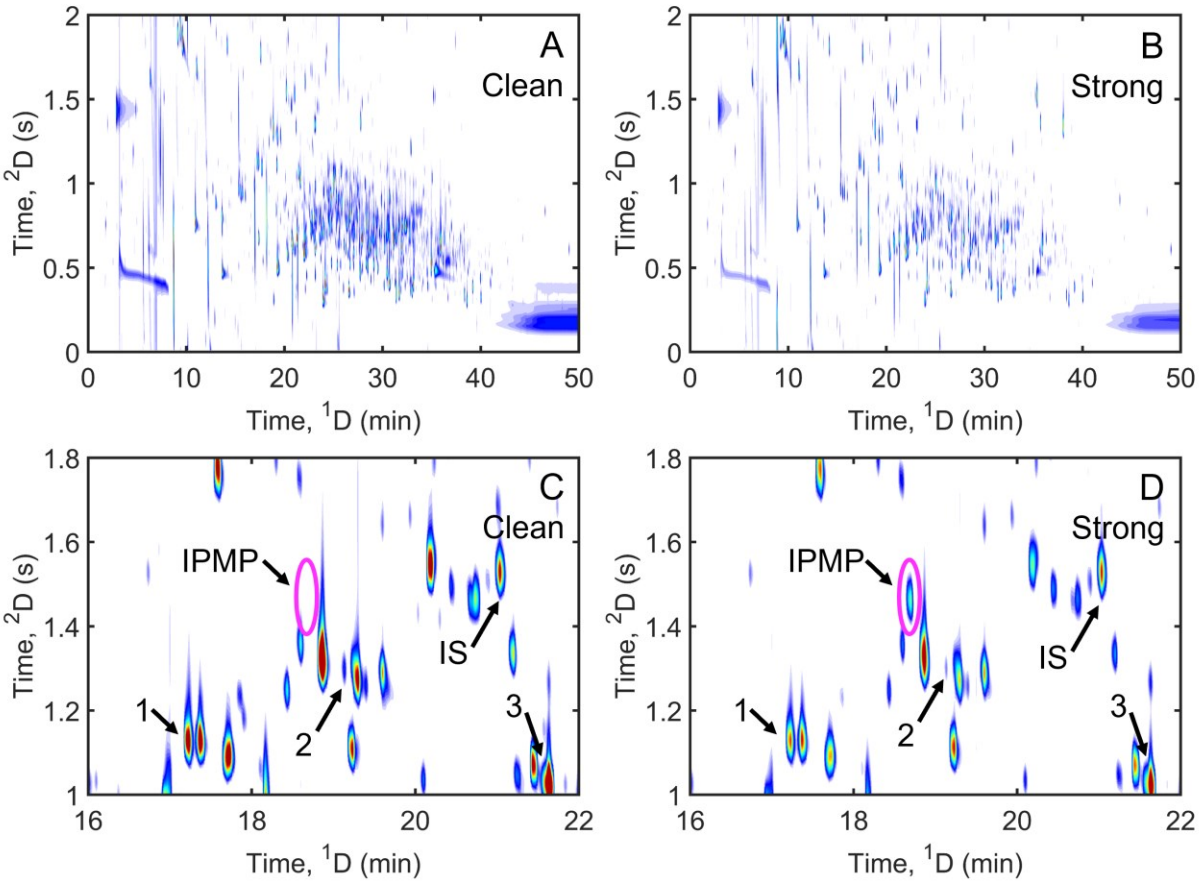


Figure 3.1. Normalized TIC GC × GC chromatograms of coffee samples categorized as clean (A) or strong PTD (B) based on their odor. Both chromatograms are plotted on the same color scale. (C-D) A zoom-in on the chromatograms from 16 to 22 min on ^{1}D and 1–1.8 s on ^{2}D.

To ensure consistency between the work presented herein and our previous 1D-GC-MS study [10], the normalized signal for IPMP in the GC×GC-TOFMS chromatograms (Figure 3.1)

collected for every coffee sample was quantified using *m/z* 152, the molecular ion for IPMP. The signals for the two replicates were then averaged together and plotted against the IPMP concentration determined with 1D-GC-MS (Figure 3.2A) [10]. The data points in Figure 3.2A are color coded according to their sensory classification: clean (red), mild PTD (yellow), medium PTD (green), and strong PTD (blue). Despite a few outlier samples, the GC×GC-TOFMS and 1D-GC-MS measurements for IPMP are in good agreement. Figure 3.2B relates the GC×GC-TOFMS intensity of IPMP to the different sensory classifications of PTD. With a *p*-value < 0.01, a one-way analysis of variance (ANOVA) statistical test demonstrates that the intensity of IPMP is statistically significantly different between the four odor attributions. Again, these results are consistent with those previously obtained [10]. Based on the results in Figure 3.2, a supervised chemometric comparison of the GC×GC-TOFMS chromatograms obtained from the clean and strong PTD coffee samples is the most promising approach for identifying analytes that contribute to the volatile fingerprint of PTD. Note, PCA was first applied to determine if the samples naturally cluster based on the presence or severity of PTD; however, no distinct clustering could be observed using the first few principal components (Figure B.2). Therefore, tile-based F-ratio analysis of the clean and strong PTD samples (the two extremes of the sensory panel classification) was selected to elucidate the PTD-related differences in the headspace of roasted coffee.

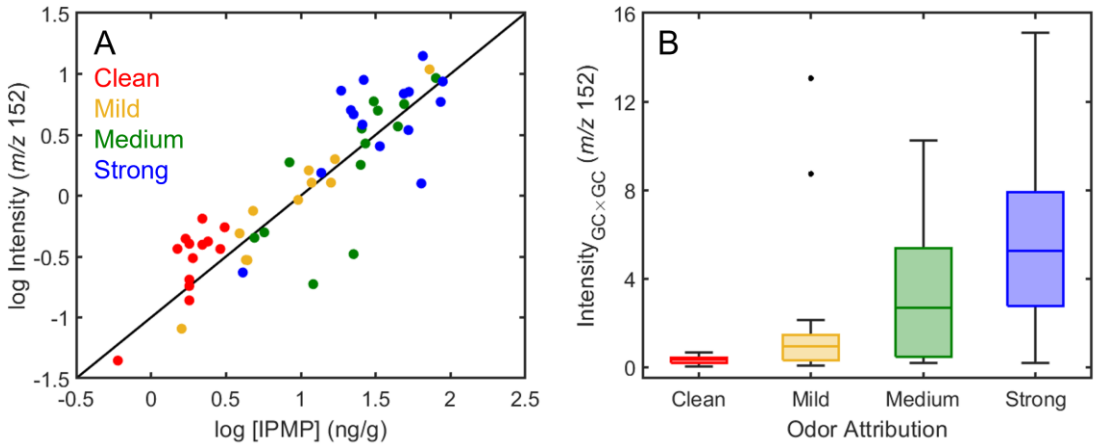


Figure 3.2. (A) Relationship between the average peak intensity of IPMP measured herein using GC×GC-TOFMS and the IPMP concentration determined in a previous 1D-GC–MS study [10]. Samples are color coded based on their sensory classification: clean (red), mild PTD (yellow), medium PTD (green), and strong PTD (blue). (B) Box-and-whiskers plot relating the peak intensity of IPMP measured herein to the different PTD odor attributions.

For tile-based F-ratio analysis, selection of the appropriate ^{1}D and ^{2}D tile dimensions is crucial to mitigate the discovery of false positives (i.e., non-class-distinguishing analytes) due to the variance produced from retention time shifting [51]. Ideally, the tile size should be large enough to encompass both the typical width of a peak and the degree of retention time shifting present between samples. Figure 3.3A-B shows the summed ^{1}D and ^{2}D peak profiles of IPMP in the clean (red) and strong PTD (blue) classes. Minor run-to-run shifting of 1 modulation (2 s) is observed as irregular peak shape in the ^{1}D peak profiles while the ^{2}D profiles show no significant signs of shifting. Therefore, as illustrated by the vertical dashed lines in Figure 3.3A-B, a tile size of 10 s × 200 ms (^{1}D × ^{2}D) was selected. Using this tile size, F-ratio analysis was performed by comparing the clean and strong PTD chromatograms. Figure 3.3C shows the F-ratio distribution for the 495 hits initially discovered, with F-ratios ranging from 65.2 to 0.01. As indicated by the arrow, IPMP was discovered near the top of the hit list (Hit #2) with an F-ratio of 54.3. Since an F-ratio was calculated for every tile possessing a summed signal greater than the *S/N* of 10 threshold, the initial hit list encompasses both class-distinguishing true positive and false positive

hits. Typically, hits discovered near the top of the hit list (larger F-ratios) are more likely to be class-distinguishing, with a larger between-class variance relative to the within-class variance, while hits at the bottom of the list (smaller F-ratios) are more likely to be false positives.

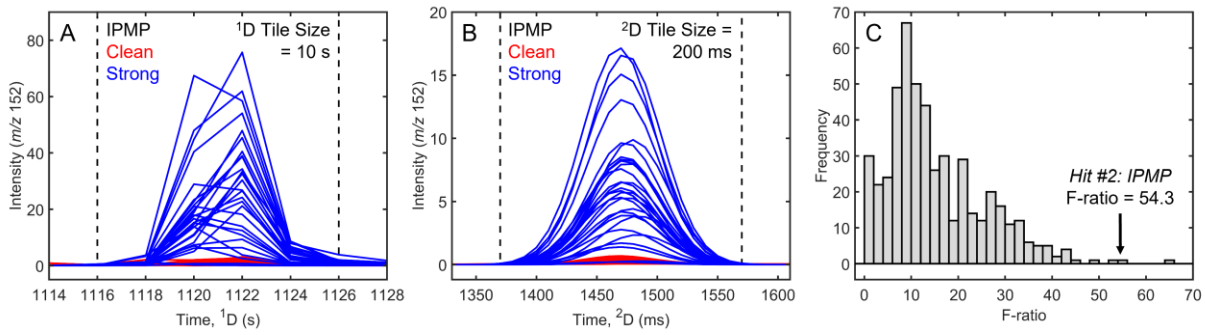


Figure 3.3. (A) The ^{1}D peak profile of IPMP at *m/z* 152 in for the clean (red) and strong PTD (blue) coffee samples. The dashed lines represent the ^{1}D tile size of 10 s. (B) The ^{2}D peak profile of IPMP at *m/z* 152 with dashed lines representing the ^{2}D tile size of 200 ms. (C) The F-ratio distribution for all 495 hits discovered. The arrow indicates the hit number and F-ratio for IPMP.

Identifying, quantifying, and determining the statistical significance for all 495 hits discovered by F-ratio analysis using a manual top-down mining approach can be burdensome. To focus identification and quantitation efforts to solely class-distinguishing analytes, a *p*-value for all 495 hits discovered by F-ratio analysis. A *p*-value from a *t*-test (assuming unequal variances) was calculated using the signal encapsulated within a 10 s × 200 ms tile surrounding each hit at the top F-ratio *m/z*. Hits with a *p*-value < 0.01 (orange) are identified as class-distinguishing (i.e., true positive) while hits with a *p*-value > 0.01 (purple) are identified as a false positive (Figure 3.4A). This *p*-value threshold, corresponding to the 99 % confidence level, was selected to minimize the erroneous inclusion of false positive hits during later identification and quantitation efforts [33]. In total, 359 out of 495 hits were determined to be class-distinguishing (*p*-value < 0.01) and became the focus for later data analysis efforts. Furthermore, the arrows on Figure 3.4A indicate the location of the first false positive (Hit #121; F-ratio = 21.1) and last true positive hit identified in the hit list (Hit #491; F-ratio = 0.14). Hence, 239 true positives were

interspersed with 136 false positives at F-ratios below 21.1, which could have been missed if the F-ratio hit list was cut-off at a certain number of hits or at a pre-determined F-ratio threshold.

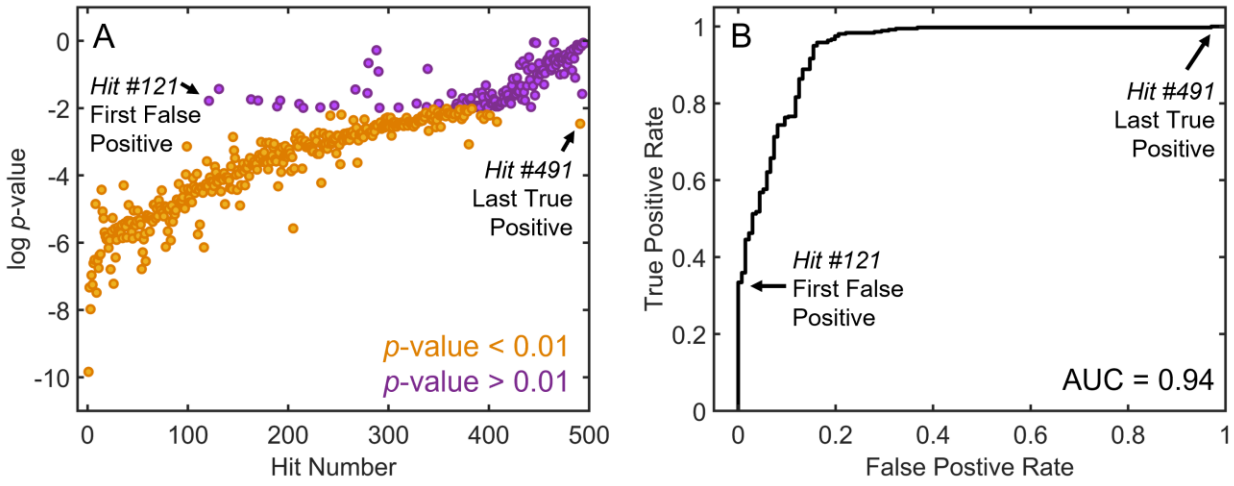


Figure 3.4. Reduction of the F-ratio hit list by determining a *p*-value threshold. (A) The *p*-value calculated for each hit using their top F-ratio *m/z*. A total of 359 hits (orange) were determined to be true positives (i.e., class-distinguishing) since their *p*-value < 0.01, which was the *p*-value threshold. The remaining 136 hits (purple) were determined to be false positives (i.e., not class-distinguishing) since their *p*-value > 0.01. The black arrows denote the first false positive (Hit #121) and last true positive (Hit #491). (B) A receiver operating characteristic (ROC) curve prepared using the results shown in (A). The first false positive (Hit #121) and last true positive (Hit #491) are denoted again for reference. The area under the curve (AUC) is also provided.

The receiver operating characteristic (ROC) curve shown in Figure 3.4B is generated for the hit list using the labels of true or false positive, defined in Figure 3.4A. ROC curves can be beneficial in chemometrics for optimizing specific parameters within a method [51,52] or comparing the performance of different chemometric methods [39,40,53]. For this study, the ROC curve highlights the importance of using a *p*-value threshold to discover analytes related to the occurrence of PTD in roasted arabica coffee. To define, a ROC curve shows the relationship between the true positive rate versus the false positive rate for a given analytical method [54]. Moving down the hit list, the true positive rate was calculated as the cumulative sum of true positive hits divided by the total number of true positives (i.e., the 359 hits with a *p*-value <

0.01). The false positive rate was calculated in a similar fashion by keeping track of the running sum and total number of false positive hits discovered. As shown in Figure 3.4B, the steps between the first false positive and last true positive hit (denoted with arrows) demonstrate how most of the class-distinguishing analytes before the last true positive hit were intermingled with a small number of false positives at low F-ratios. Additionally, the area under the ROC curve (AUC) defines the probability that the *p*-value threshold correctly distinguished between true or false positives, where an AUC of 1 represents the maximum classifying power [54]. The AUC for the ROC curve in Figure 3.4B equals 0.93, which means that the *p*-value threshold of 0.01 could distinguish between significant chemical differences and background variation with high accuracy. Ultimately, the high AUC speaks to the outstanding performance of the tile-based F-ratio software [55].

The first 30 identifiable class-distinguishing (i.e., true positive) analytes discovered by F-ratio analysis are listed in Table 3.1, while Table B.2 provides similar information for all 359 true positives identified in Figure 3.4. A MV $\geq$ 800 between the hit and library mass spectrum was required for tentative analyte identification [45]. However, identification for some of these class-distinguishing analytes was challenging due to the presence of larger, overlapping interferent signals, so a chemometric decomposition method known as MCR-ALS was applied in these situations. In total, 145 analytes could not be identified even after MCR-ALS decomposition (Table B.2). This work discovered many analytes that were previously identified in the literature as a potential marker for PTD [10]. However, discrepancies between this work and previous literature [10,11] may exist due to differences in the roasting degree of the coffee beans or chromatographic methodology. Table 3.1 also highlights several analytes that have not been documented in coffee beans previously, denoted by a dagger (†). Note, these compound

names are tentative since the identifications are based off the analyte mass spectrum alone; future studies with chemical standards can confirm their identification. However, many of these tentative compounds are structurally like other volatiles documented in coffee [2,3,56–60]. The pure analyte *m/z* used for quantifying the concentration ratio ([Strong]/[Clean]) is also provided along with its *p*-value and *LOF*. Twenty-three out of the 359 total analytes were only present in one class (marked with an asterisk; Table 3.1 & Table B.2). The [Strong]/[Clean] for analytes that had a larger signal in the clean samples ranged from 0.02 – 0.82 while the [Strong]/[Clean] for analytes with a larger signal in the strong PTD samples ranged from 1.36 – 29.81 (Table B.2). Since many volatiles in the coffee headspace are responsible for the final aroma of the beans, Table 1 and S2 reports the known organoleptic / odor properties [50] for 96 of the discovered analytes. Closer examination of these sensory descriptions highlights that analytes with desirable coffee aromas (e.g., roasted, nutty, cocoa, fruity) were found in decreased abundance in the strong PTD samples. Conversely, analytes discovered with larger signals in the strong PTD samples had less favorable aromas (e.g., vegetable-like, musty, aldehydic). These results further support the hypothesis that the diminished concentration of compounds with pleasant aromas further amplifies the odor attributed to PTD [10].

Table 3.1. The first 30 identifiable hits (MV ≥ 800) which were discovered by F-ratio analysis. Compounds not previously identified in coffee are denoted by a dagger (†). A concentration ratio for each analyte was calculated as [Strong]/[Clean] ([S]/[C]) using a pure *m/z* based upon applying the S-ratio algorithm [47]. The metrics for determining *m/z* purity (*p*-value and *LOF*) are also reported. Analytes present in only one sample class are denoted by an asterisk (*). For these analytes, only a *p*-value is reported. Sensory descriptions are listed for known analytes [50].

Hit Number	Time, ^{1}D (min)	Time, ^{2}D (s)	Compound	MV	S-ratio *m/z*	[S]/[C]	*p*-value	*LOF* (%)	Sensory Description
1 †	26.57	0.77	7-Benzofuranamine, 2-methyl-	811	147	0.55	7.5E-11	15.0	
2	18.67	1.47	IPMP	849	152	20.4	4.9E-08	8.60	Earthy, Vegetable, Potato
4 †	38.30	1.16	1,3-Pentadiene, 1,1-diphenyl-, (*Z*)-	823	205	10.9	9.7E-08	15.7	
5 †	25.03	0.86	3(2H)-Benzofuranone, 7-methyl-	801	148	0.47	4.7E-08	10.3	
6 †	39.77	1.43	Benzene, 1,1'-(1,1,2,2-tetramethyl-1,2-ethanediyl)bis-	817	119	20.7	1.8E-07	5.60	
7 †,*	37.23	1.60	Benzene, 1,1'-(1,4-dimethyl-1-butene-1,4-diyl)bis-	809	221	Strong only	3.1E-07		
8 †	37.83	1.29	1,5,6,7-Tetramethyl-3-phenylbicyclo[3.2.0]hepta-2,6-diene	803	194	15.0	2.6E-05	19.7	
9	17.07	0.33	Pyridine, 3-ethyl-	813	136	0.31	5E-08	18.2	Caramellic, Roasted, Hazelnut
10	22.80	0.94	Benzofuran, 2-methyl-	826	103	0.42	5.3E-07	12.7	Burnt, Phenolic
12	17.67	0.27	2,4,6-Octatriene, 2,6-dimethyl-	855	79	0.46	7.8E-09	15.6	Sweet, Floral, Nutty
13 †	23.70	0.87	1H-Indole, 2,3-dihydro-	801	117	0.49	1.9E-07	15.6	
16	22.63	0.93	2-Propenal, 3-phenyl-	869	133	0.17	5.9E-08	7.60	Sweet, Spicy, Honey, Cinnamon
17	14.40	1.02	Cyclohexene, 1-methyl-4-(1-methylethylidene)-	902	103	0.81	3.1E-04	16.8	Fresh, Sweet, Pine, Citrus
21	21.07	1.20	3-Methyl-2,3-dihydro-benzofuran	803	105	0.39	2.5E-07	14.2	
24 †	35.77	1.37	1-Propene, 3-(2-cyclopentenyl)-2-methyl-1,1-diphenyl-	810	222	29.8	1.9E-06	10.8	
25	17.97	1.34	2-Methyl-3-isopropylpyrazine	820	108	0.43	7E-07	10.7	Coffee
28	24.70	0.86	Furan, 2-(2-furanylmethyl)-5-methyl-	910	74	0.46	2.9E-07	9.10	
31	13.50	0.79	1,3,7-Octatriene, 3,7-dimethyl-	924	93	0.32	2.6E-06	14.0	Fruity, Floral
32	27.60	0.85	Phenylethyl acetate	802	159	0.53	1.1E-06	10.0	Floral, Sweet, Honey, Fruity, Cocoa
33	17.43	1.51	Pyrazine, 2-ethyl-5-methyl-	877	93	0.14	1.1E-07	11.6	Coffee, Nutty, Roasted
34	37.23	0.66	2,4-Di-*tert*-butylphenol	908	191	4.50	2.5E-06	4.00	

36	21.40	1.52	*cis*-4-Decenal	809	98	0.41	3.5E-06	5.10	Citrus, Aldehydic, Cardamom
37 †	35.40	1.53	1,1,3-Trimethyl-3-phenylindan	896	236	19.6	6.8E-06	4.80	
38	21.83	1.03	5,6,7,8-Tetrahydroquinoxaline	805	119	0.50	6.7E-07	19.5	Nutty, Roasted, Cereal
39 †	24.97	1.09	Benzofuran, 4,7-dimethyl-	862	144	0.57	2.9E-08	5.60	
40 †	38.90	1.31	2,4-Diphenyl-4-methyl-2(*E*)-pentene	883	236	19.8	4.3E-06	4.90	
41	23.03	0.82	2-Furfurylfuran	901	100	0.47	6.8E-06	14.8	Rich, Roasted
42	17.83	1.24	3-Octen-2-one, (*E*)-	874	68	0.49	9.2E-07	18.4	
43 †,*	37.23	1.23	Benzene, (1,3-dimethyl-3-butenyl)-	807	105	Strong only	3.9E-06		
45 †,*	39.33	1.44	Benzene, [2-methyl-1-(1-methylethyl)propyl]-	801	91	Strong only	7.8E-07		

The locations of all 359 class-distinguishing analytes (Table 3.1 & Table B.2) can be visualized as stitch GC×GC chromatograms, which are shown in Figure 3.5A-B. These stitch chromatograms categorize the analytes as either having higher signal in the clean (A) or strong PTD (B) coffee samples. This grouping was based on their determined concentration ratio ([Strong]/[Clean]) using the S-ratio algorithm [47]. Based on Figure 3.5A-B, 327 analytes discovered had statistically higher abundance in the clean samples ([Strong]/[Clean] < 0.82), whereas the remaining 32 analytes had higher signals in the strong PTD class ([Strong]/[Clean] > 1.36). These stitch chromatograms can also be visualized by projecting the determined concentration ratios onto windows surrounding every peak (Figure 3.5C). As referenced earlier, the TIC chromatograms of a clean and strong PTD sample highlighted an overall decrease in signal for the strong PTD samples (Figure 3.1). The results in Figure 3.5 demonstrate that this overall lower signal in the TIC chromatogram is due to a decrease in analyte concentrations in the strong PTD samples. Additionally, visual comparison of the TIC (Figure 3.1) and stitch (Figure 3.5) chromatograms demonstrates that F-ratio analysis discovered many class-distinguishing analytes, which were not observed in Figure 3.1. These analytes were not observed in the TIC chromatogram due to their signal being smaller than the noise when all the

m/z were summed together [48]. This result illustrates that both F-ratio analysis can discover analytes near the limit of quantitation [61] and PTD is responsible for altering the volatile profile of coffee by affecting analytes at all concentration levels.

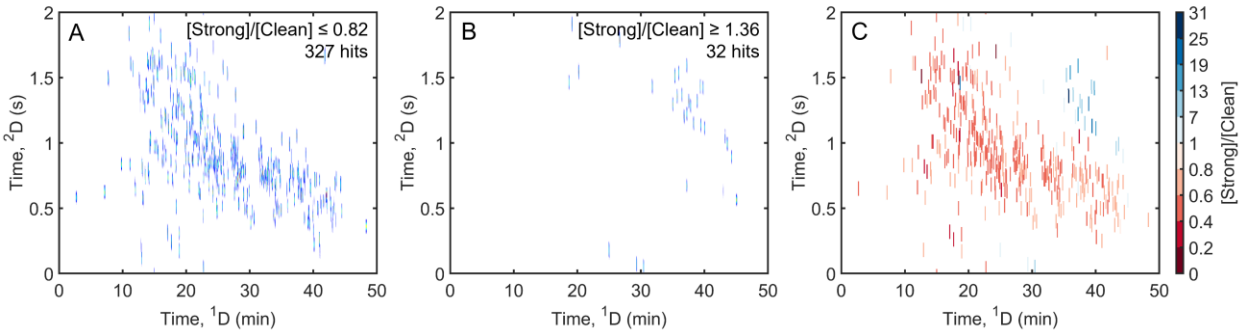


Figure 3.5. Visualization of the 359 class-distinguishing hits discovered by F-ratio analysis. (A) Stitch GC×GC chromatogram of the 327 hits that were discovered to have a higher signal in clean coffee samples ([Strong]/[Clean] ≤ 0.82). For each hit, the sample with maximum signal at the S-ratio *m/z* [47] was extracted from the data and placed into the stitch chromatogram. (B) Stitch GC×GC chromatogram of the 32 hits that were discovered to have a higher signal in strong PTD coffee samples ([Strong]/[Clean] ≥ 1.36). (C) Projection of the calculated concentration ratios on the window surrounding each peak shown in (A-B).

The resulting PCA scores (A) and loadings (B) plots shown in Figure 3.6 after reducing the chromatographic data down to the signals for the 359 analytes discovered by F-ratio analysis, and then found to significantly change in concentration between the clean and strong PTD sample classes. Compared to the original PCA model (Figure B.2), which captured only 53.6 % of the variance, the PCA model shown in Figure 3.6A-B now captures 91.79 % of the total variance within the data. The increased percent variance captured is due to the reduction of background noise and retention time misalignment in the data [62]. The scores plot in Figure 3.6A also now illustrates that the coffee samples cluster based on the presence of PTD, where the clean samples (red) are separated from those affected by PTD at any odor attribution level (yellow, green, and blue). Note, there is no observable difference in the PCA model after excluding the signal of IPMP from the data set (Figure B.3), so the sample clustering observed

shown in Figure 3.6 was not solely due to IPMP. The differentiation between the clean samples and those affected by PTD primarily occurs along PC 1 (Figure 3.6A). Inspection of the PCA loadings (Figure 3.6B) emphasizes that this differentiation along PC 1 is driven by five class-distinguishing analytes: IPMP (Hit #2), 3-(2-cyclopentenyl)-2-methyl-1,1-diphenyl-1-propene (Hit #24), 2,4-di-*tert*-butylphenol (Hit #34), 1,1,3-trimethyl-3-phenylindan (Hit #37), and 2,4-diphenyl-4-methyl-2(*E*)-pentene (Hit #40). All five of these highly loaded analytes were found to have large signal differences between the clean and strong PTD class, with [Strong]/[Clean] ranging from 4.5 to 29.8 (Table 3.1). Figure 3.6B also demonstrates that most of the analytes discovered have more subtle differences between their classes since their loadings cluster around zero. However, these analytes with minor concentration differences are still pertinent since they contribute to the aroma profile of coffee.

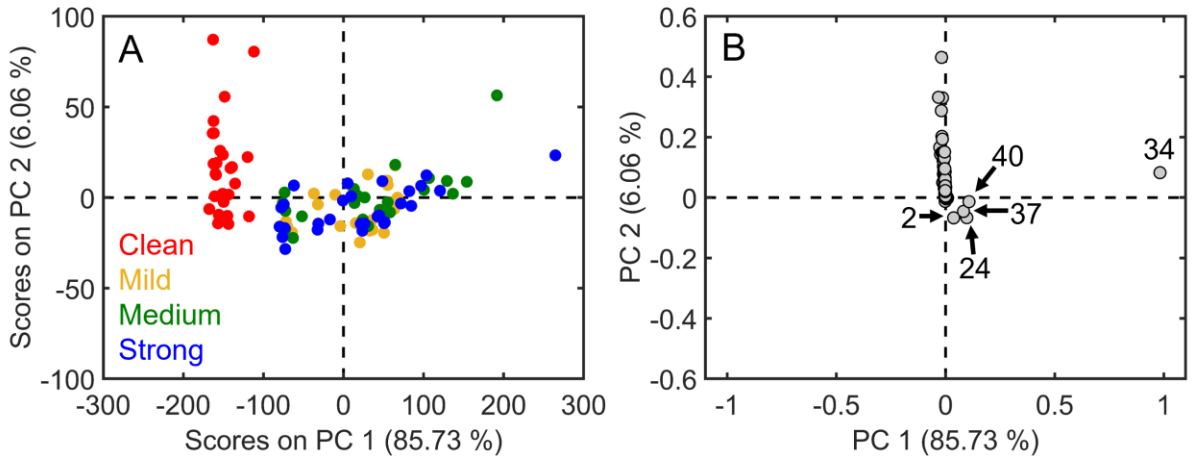


Figure 3.6. Results from PCA using the normalized intensity measured at the S-ratio *m/z* [47] for the discovered hits. (A) Scores plot for the model built using the signal for all 359 statistically significant hits in the clean (red), mild PTD (yellow), medium PTD (green), and strong PTD (blue) samples. (B) Loadings plot for the model shown in (A), where each gray dot corresponds to one of the statistically significant hits discovered. Five highly loaded hits (Hit #2, 24, 34, 37, and 40) on PC 1 are labeled.

The results presented in Figure 3.6 and previous literature [4,8–11] highlight that IPMP and a small handful of analytes are responsible for differentiating clean and PTD-affected coffee

beans. However, this current study aims to delve deeper and showcase how a complete volatile profile of these roasted coffee beans can contribute to a fuller understanding of PTD. For this goal, PLS regression was selected to relate the compounds discovered by F-ratio analysis to the concentration of IPMP. Figure 3.7 shows the PLS prediction of IPMP concentration using the entire F-ratio hit list, excluding IPMP (Hit #2) since it was being predicted. The regression plot in Figure 3.7A highlights the relationship between the IPMP concentration measured previously [10] and the concentration predicted by the PLS model, where each sample is color coded according to its sensory information. While the data in Figure 3.7A is color coded according to the odor attributed to PTD, it is important to note that the PLS regression does not take this sensory information into account when developing a model. Ideally, the measured and predicted concentrations should be equal (i.e., fall along the black 1:1 line). Figure 3.7A shows that the PLS model developed using the F-ratio hit list can accurately predict IPMP concentration since the samples cluster closely around the 1:1 line and the model has low prediction errors (< 11 %). Using this PLS model, the linear regression vector (LRV) can be investigated to determine how each volatile analyte discovered by F-ratio analysis correlates with the concentration of IPMP. Note, discovering the relationship that each analyte has with IPMP is a direct benefit of coupling PLS modeling with tile-based F-ratio analysis. The LRV value for each analyte is provided in Table B.2. Figure 3.7B displays the LRV as a GC×GC chromatogram, where the value for each hit is projected onto a 2D window surrounding its retention time location. For the PLS model, a positive LRV (blue) indicates that the given compound has a direct relationship with IPMP concentration while a negative LRV (red) indicates an inverse relationship between the compound and IPMP. A visual comparison of the LRV in Figure 3.7B and the projection of the concentration ratio for each analyte in Figure 3.5C shows a striking similarity, where every

analyte with a [Strong]/[Clean] > 1.36 has a positive LRV value and every analyte with a [Strong]/[Clean] < 0.82 has a negative LRV value in the PLS model. Table 3.2 and Table 3.3 provide a list of the top 20 compounds identified in the LRV with a positive and negative loading, respectively. Despite the variation observed in each sensory class (Figure 3.2), the PLS model accurately utilizes the entire volatile profile of PTD to predict IPMP concentration. Ultimately, the PLS modeling results underscore that the entire coffee headspace, not just IPMP and a select number of analytes, plays a significant role in PTD.

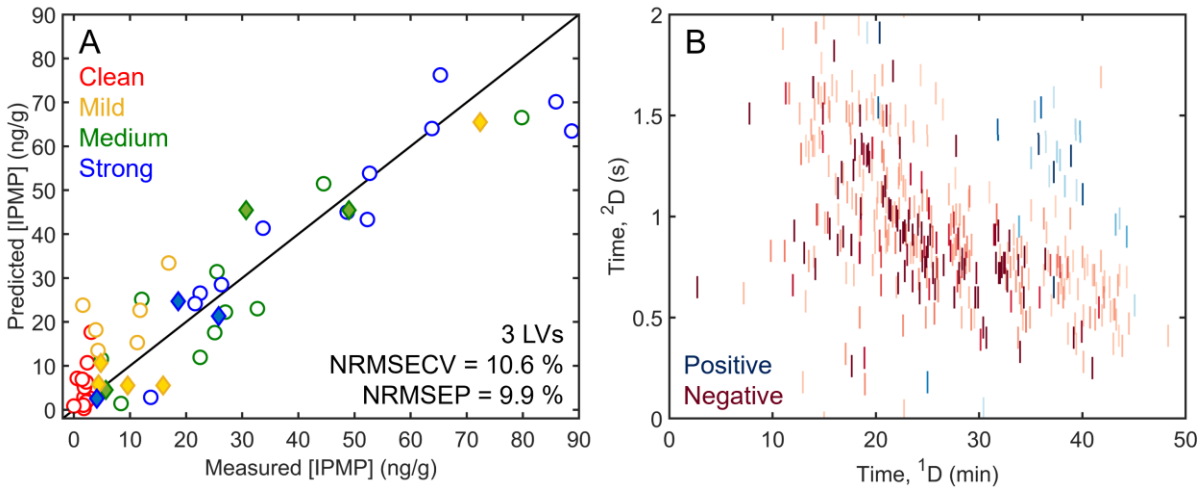


Figure 3.7. PLS prediction of IPMP concentration using the normalized intensity measured at the S-ratio *m/z* [47] for all discovered hits except for IPMP (Hit #2). (A) Regression plot for the PLS model. The black line symbolizes ideal agreement between the predicted and measured concentrations. Samples used to build the calibration model are shown as unfilled circles while samples used in the external validation set are shown as filled diamonds. Samples are color coded based on their sensory classification: clean (red), mild PTD (yellow), medium PTD (green), and strong PTD (blue). The number of LVs, NRMSECV, and NRMSEP for each PLS model is provided. (B) Projection of the linear regression vector value for each peak on its surrounding window. Positive loadings are highlighted in blue while negative loadings are highlighted in red.

Table 3.2. The top 20 discovered hits with a positive loading in the LRV of the PLS model. The hit list is ranked in descending order of their F-ratio hit number. S-ratios for each analyte were calculated as [Strong]/[Clean] using a pure *m/z*. Tentative compound identifications were made if the mass spectrum match a library spectrum with a MV ≥ 800. Peaks that could not be identified are listed as an unknown (Unk) and numbered according to their order in the hit list (Table B.2).

LRV Ranking	LRV	F-ratio Hit Number	Time, ^{1}D (min)	Time, ^{2}D (s)	Compound	MV	[Strong]/[Clean]
1	1.16E-03	34	37.23	0.66	2,4-Di-*tert*-butylphenol	908	4.50
2	8.52E-05	40	38.90	1.31	2,4-Diphenyl-4-methyl-2(*E*)-pentene	883	19.8
3	5.85E-05	116	20.37	1.92	Decanal	805	1.63
4	1.34E-05	24	35.77	1.37	1-Propene, 3-(2-cyclopentenyl)-2-methyl-1,1-diphenyl-	810	29.8
5	1.14E-05	43	37.23	1.23	Benzene, (1,3-dimethyl-3-butenyl)-	807	Strong only
6	8.89E-06	20	40.03	1.12	Unk7		Strong only
7	8.23E-06	149	20.23	1.55	Unk50		2.70
8	7.76E-06	11	31.80	1.44	Unk2		5.17
9	7.51E-06	155	36.00	1.55	Propane, 2-cyclohexyl-2-phenyl-	801	Strong only
10	7.06E-06	99	25.00	0.19	2-Undecanone, 6,10-dimethyl-	923	1.36
11	5.14E-06	282	44.30	0.91	Unk92		4.86
12	4.92E-06	6	39.77	1.43	Benzene, 1,1'-(1,1,2,2-tetramethyl-1,2-ethanediyl)bis-	817	20.7
13	4.80E-06	3	43.23	0.97	Unk1		2.81
14	4.22E-06	252	33.87	0.96	Acenaphthene	901	2.04
15	3.11E-06	325	35.03	1.30	1,1'-Biphenyl, 3,4-diethyl-	812	1.43
16	3.07E-06	37	35.40	1.53	1,1,3-Trimethyl-3-phenylindan	896	19.6
17	3.07E-06	8	37.83	1.29	1,5,6,7-Tetramethyl-3-phenylbicyclo[3.2.0]hepta-2,6-diene	803	15.1
18	2.93E-06	19	37.53	1.33	Unk6		Strong only
19	2.86E-06	4	38.30	1.16	1,3-Pentadiene, 1,1-diphenyl-, (*Z*)-	823	10.9
20	2.41E-06	364	45.07	0.57	1,4-Benzenediol, 2,6-bis(1,1-dimethylethyl)-	802	1.38

Table 3.3. The top 20 discovered hits with a negative loading in the LRV of the PLS model. The hit list is ranked in descending order of their F-ratio hit number. S-ratios for each analyte were calculated as [Strong]/[Clean] using a pure *m/z*. Tentative compound identifications were made if the mass spectrum match a library spectrum with a MV ≥ 800. Peaks that could not be identified are listed as an unknown (Unk) and numbered according to their order in the hit list (Table B.2).

LRV Ranking	LRV	F-ratio Hit Number	Time, ^{1}D (min)	Time, ^{2}D (s)	Compound	MV	[Strong]/[Clean]
1	-1.29E-03	210	15.53	0.95	Pyrazine, 2,6-dimethyl-	915	0.56
2	-8.74E-04	153	28.37	0.80	2-Naphthalenol	886	0.60
3	-8.74E-04	306	30.70	0.41	3-Acetylpyrrole	869	0.60
4	-4.34E-04	261	32.40	0.78	Unk84		0.63
5	-4.00E-04	271	22.17	0.68	Furyl ethyl ketone	874	0.59
6	-2.02E-04	200	31.63	0.78	2,7-Naphthalenediol	859	0.55
7	-1.91E-04	120	24.07	0.97	Pyrazine, 2-methyl-5-(1-propenyl)-, (*Z*)-	875	0.50
8	-8.29E-05	148	19.37	1.28	Pyrazine, 2-methyl-6-propyl-	845	0.31
9	-7.20E-05	258	32.20	0.76	Unk83		0.57
10	-6.20E-05	5	25.03	0.86	3(2H)-Benzofuranone, 7-methyl-	801	0.47
11	-5.70E-05	81	24.23	0.63	2,2'-Bifuran	801	0.45
12	-5.55E-05	217	22.37	0.89	3-Acetyl-2,5-dimethyl furan	863	0.52
13	-5.48E-05	117	25.73	0.65	3-Methyl-2-thiophenecarboxaldehyde	901	0.52
14	-5.46E-05	51	28.57	0.75	4-Hydroxybenzo[b]thiophene	809	0.54
15	-5.35E-05	113	24.77	0.76	2-Acetyl-3-methylpyrazine	863	0.10
16	-4.10E-05	255	34.40	0.81	Unk81		0.48
17	-4.08E-05	1	26.57	0.77	7-Benzofuranamine, 2-methyl-	811	0.55
18	-4.06E-05	223	29.63	0.61	2-Thiophenecarboxylic acid, 4-nitrophenyl ester	805	0.54
19	-3.88E-05	245	33.80	0.77	Thiophene, 2-phenyl-	890	0.29
20	-3.74E-05	10	22.80	0.94	Benzofuran, 2-methyl-	826	0.42

Box-and-whiskers plots relating the normalized intensity for eight exemplary analytes to their PTD odor attribution are shown in Figure 3.8. Additional compounds are shown in Figure B.4. The top row (Figure 3.8A-D) focuses on four analytes that were responsible for the sample clustering on the PCA scores plot (Figure 3.6) and had positive loading in the PLS model (Figure 3.7; Table 3.2): 3-(2-cyclopentenyl)-2-methyl-1,1-diphenyl-1-propene, 2,4-diphenyl-4-methyl-

2(*E*)-pentene, 2,4-di-*tert*-butylphenol, and 1,1,3-trimethyl-3-phenylindan. Many of the volatiles highlighted in Figure 3.8A-D and Table 3.2 are aromatic hydrocarbons and oxygenated compounds, with their signals elevated in the PTD-affected coffee samples. The presence of these analytes can potentially elucidate the biochemical mechanism linking antestia bug damage to PTD. For example, one potential pathway is that bug damage to the coffee plant creates favorable conditions for microorganisms, with research highlighting that the bacteria and fungi found on PTD-affected coffee beans produced IPMP as a metabolite [12–14]. Previous research has shown that the presence of microorganisms on coffee beans can cause the concentration of various compound classes like hydrocarbons, phenols, ketones, and aldehydes in coffee to increase [3,63,64]. These types of compounds can form by microorganisms oxidizing the lipids naturally present in the coffee beans [3]. Lipid oxidation induced by microorganisms can potentially explain the increased signals observed in Figure 3.8A-B and other analytes in Table 3.2 like 6,10-dimethyl-2-undecanone (Hit #99) and decanal (Hit #116). Phenols and phenylindanes like the analytes highlighted in Figure 3.8C-D are formed during the roasting process via the degradation of chlorogenic acids [3,58]. Studies on defective Brazilian coffee beans found higher levels of chlorogenic acids in microbe affected green coffee [3,65]. Hence, the increased abundance of 2,4-di-tert-butylphenol and 1,1,3-trimethyl-3-phenylindan could be due to the PTD-affected coffee beans having a higher concentration of chlorogenic acids prior to roasting. More specifically, 2,4-di-*tert*-butylphenol has also been identified as a volatile indicative of bacterial growth on food products [66,67]. Thus, these results highlight that microbe damage could be a potential cause of PTD.

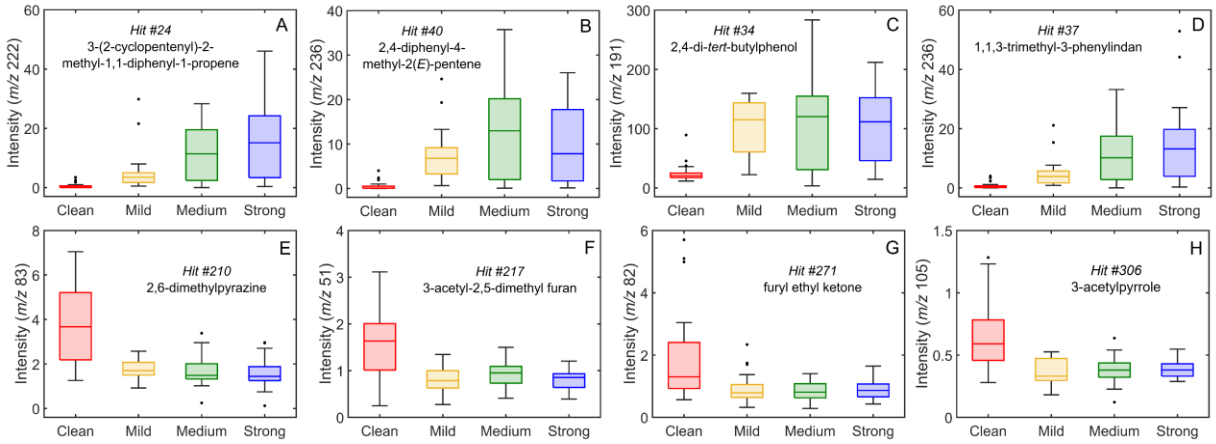


Figure 3.8. Box-and-whiskers plots relating the intensity measured at the S-ratio m/z [47] to their PTD odor attribution for eight analytes that were highly loaded in the PCA and PLS models. The top row highlights analytes with signals larger in the PTD affected samples: (A) 3-(2-cyclopentenyl)-2-methyl-1,1-diphenyl-1-propene, (B) 2,4-diphenyl-4-methyl-2*(E)*-pentene, (C) 2,4-di-*tert*-butylphenol, and (D) 1,1,3-trimethyl-3-phenylindan. The bottom row highlights analytes with signals larger in the clean coffee samples: (E) 2,6-dimethylpyrazine, (F) 3-acetyl-2,5-dimethyl furan, (G) furyl ethyl ketone, and (H) 3-acetylpyrrole.

In contrast, Figure 3.8E-H shows the diminished intensity of four analytes that had a negative loading in the PLS model (Figure 3.7; Table 3.3): 2,6-dimethylpyrazine, 3-acetyl-2,5-dimethyl furan, furyl ethyl ketone, and 3-acetylpyrrole. Although these analytes were downregulated in samples affected by PTD, their signal remains invariant across the mild, medium, and strong PTD beans (Figure 3.8D-H). Pyrazines and furans like those volatiles highlighted in Figure 3.8E-F are the most important contributors to coffee aroma [3], providing cocoa, nutty, and roasted notes [50]. Ketones and pyrroles (examples shown in Figure 3.8G-H), while not primary coffee odorants [3], can also provide sweet and fruity aromas [50]. These analytes primarily form during the roasting process, where major chemical reactions convert sugars, lipids, proteins, and chlorogenic acids into hundreds of volatile components [16]. For instance, a previous study found that sucrose and other carbohydrates were more concentrated in non-defective green coffee beans relative to those affected by microbial growth [3]. In turn, the non-defective beans had higher levels of furans and pyrroles after roasting [3]. The same study

also indicated that roasted defective coffee beans could exhibit a higher concentration of alkylpyrazines due to an imbalance in amino acid and sugar content, thereby promoting the formation of pyrazines [3]. The work herein and previously reported [10] suggests that alkylpyrazines are less concentrated in PTD-impacted samples [10]. Given the complexity of the roasting process, future work is needed to elucidate how the chemical composition in both clean coffee beans and those affected by PTD impacts the volatile profile of the ultimate roasted product. Nonetheless, the decreased abundance of these analytes suggests that their absence contributes to PTD odor severity and may be further indication that biochemical changes are occurring in PTD-affected beans prior to roasting.

3.4. Conclusion

Volatile fingerprinting of PTD in roasted arabica coffee beans was performed using HS-SPME-GC×GC-TOFMS and non-targeted chemometrics. Tile-based F-ratio analysis discovered 359 analytes that changed with statistical significance, including IPMP, that differentiated clean and strong PTD coffee samples. These analytes were quantified using a pure *m/z*, providing a determination of the concentration ratio. Most of the analytes (327 out of 359 hits) had higher signals in the clean coffee samples with analyte concentrations ranging from present only in clean class to a [Strong]/[Clean] equal to 0.82. Meanwhile, only 32 analytes had larger abundances for the strong PTD class with their [Strong]/[Clean] ranging from 1.36 to only being present in these samples. Notably, examination of the known sensory properties for some volatiles revealed that analytes with desirable coffee aromas were found in decreased abundance in the strong PTD samples. PCA using the signals for these 359 hits illustrated sample clustering based on the presence of PTD while PLS modeling demonstrated that the compounds discovered by F-ratio analysis can accurately predict the concentration of IPMP. Compounds that are

heavily weighted in both the PCA and PLS loadings implies that damage from microorganisms is one possible pathway for PTD. In turn, the damage from microorganisms can induce biochemical changes, which decrease the concentration of analytes with positive aroma descriptions. However, it is important to note that this research cannot exclude that these biochemical changes are also caused by the stress response pathway of *O*-methyltransferase expression in the coffee plant [15]. To further understand the differences in the volatile profile of samples affected by PTD, more work is required to deepen the link between antestia bug predation and the biochemical processes occurring inside green coffee beans and during roasting.

3.5. References

[1] P.D.C. Mancha Agresti, A.S. Franca, L.S. Oliveira, R. Augusti, Discrimination between defective and non-defective Brazilian coffee beans by their volatile profile, Food Chem. 106 (2008) 787–796. https://doi.org/10.1016/j.foodchem.2007.06.019.

[2] A.T. Toci, A. Farah, Volatile compounds as potential defective coffee beans' markers, Food Chem. 108 (2008) 1133–1141. https://doi.org/10.1016/j.foodchem.2007.11.064.

[3] A.T. Toci, A. Farah, Volatile fingerprint of Brazilian defective coffee seeds: Corroboration of potential marker compounds and identification of new low quality indicators, Food Chem. 153 (2014) 298–314. https://doi.org/10.1016/j.foodchem.2013.12.040.

[4] R. Becker, B. Dohla, S. Nitz, O.G. Vitzthum, Identification of the "Peasy" Off-Flavour Note in Central African Coffees, in: 12th Int. Sci. Colloq. Coffee, Montreaux, Switzerland, 29 June - 3 July 1987, Association for Science and Information on Coffee (ASIC), Paris, France, 1987: pp. 203–215.

[5] B. Bouyjou, B. Decazy, G. Fourny, Removing the "potato taste" from Burundian Arabica, Plant. Rech. Dev. 6 (1999) 107–115.

[6] A.G. Ahmed, L.K. Murungi, R. Babin, Developmental biology and demographic parameters of antestia bug Antestiopsis thunbergii (Hemiptera: Pentatomidae), on Coffea arabica (Rubiaceae) at different constant temperatures, Int. J. Trop. Insect Sci. 36 (2016) 119–127. https://doi.org/10.1017/S1742758416000072.

[7] J. Bigirimana, A. Gerard, D. Mota-Sanchez, L.J. Gut, Options for Managing Antestiopsis thunbergii (Hemiptera: Pentatomidae) and the Relationship of Bug Density to the Occurrence of Potato Taste Defect in Coffee, Florida Entomol. 101 (2018) 580–586. https://doi.org/10.1653/024.101.0418.

[8] S.C. Jackels, E.E. Marshall, A.G. Omaiye, R.L. Gianan, F.T. Lee, C.F. Jackels, GCMS investigation of volatile compounds in green coffee affected by potato taste defect and

the antestia bug, J. Agric. Food Chem. 62 (2014) 10222–10229. https://doi.org/10.1021/jf5034416.

[9] D. Mutarutwa, L. Navarini, V. Lonzarich, P. Crisafulli, D. Compagnone, P. Pittia, Determination of 3-Alkyl-2-methoxypyrazines in Green Coffee: A Study to Unravel Their Role on Coffee Quality, J. Agric. Food Chem. 68 (2020) 4743–4751. https://doi.org/10.1021/acs.jafc.9b07476.

[10] C.N. Cain, N.J. Haughn, H.J. Purcell, L.C. Marney, R.E. Synovec, C.T. Thoumsin, S.C. Jackels, K.J. Skogerboe, Analytical Determination of the Severity of Potato Taste Defect in Roasted East African Arabica Coffee, J. Agric. Food Chem. 7 (2021) 2253-2261. https://doi.org/10.1021/acs.jafc.1c00605.

[11] J.B. Shingiro, P.K. Shee, R.M. Beaudry, D. Thiagarajan, L.D. Bourquin, K.D. Walker, Assessing Alkyl Methoxypyrazines as Predictors of the Potato-Taste Defect in Coffee, ACS Food Sci. Technol. 2 (2022) 1738–1745. https://doi.org/10.1021/acsfoodscitech.2c00233.

[12] D. Gueule, G. Fourny, E. Ageron, A. Le Flèche-Matéos, M. Vandenbogaert, P.A.D. Grimont, C. Cilas, Pantoea coffeiphila sp. nov., cause of the 'potato taste' of Arabica coffee from the African great lakes region, Int. J. Syst. Evol. Microbiol. 65 (2015) 23–29. https://doi.org/10.1099/ijs.0.063545-0.

[13] J.B. Ndayambaje, A. Nsabimana, S. Dushime, F. Ishimwe, H. Janvier, M.P. Ongol, Microbial identification of potato taste defect from coffee beans, Food Sci. Nutr. 7 (2019) 287–292. https://doi.org/10.1002/fsn3.887.

[14] A.R. Hale, P.M. Ruegger, P. Rolshausen, J. Borneman, J. in Yang, Fungi associated with the potato taste defect in coffee beans from Rwanda, Bot. Stud. 63 (2022). https://doi.org/10.1186/s40529-022-00346-9.

[15] K.E. Frato, Identification of Hydroxypyrazine O-Methyltransferase Genes in Coffea arabica: A Potential Source of Methoxypyrazines That Cause Potato Taste Defect, J. Agric. Food Chem. 67 (2019) 341–351. https://doi.org/10.1021/acs.jafc.8b04541.

[16] W.B. Sunarharum, D.J. Williams, H.E. Smyth, Complexity of coffee flavor: A compositional and sensory perspective, Food Res. Int. 62 (2014) 315–325. https://doi.org/10.1016/j.foodres.2014.02.030.

[17] J.M. Davis, J.C. Giddings, Statistical theory of component overlap in multicomponent chromatograms, Anal. Chem. 55 (1983) 418–424. https://doi.org/10.1021/ac00254a003.

[18] Z. Liu, J.B. Phillips, Comprehensive two-dimensional gas chromatography using an on-column thermal modulator interface, J. Chromatogr. Sci. 29 (1991) 227–231. https://doi.org/10.1093/chromsci/29.6.227.

[19] M.S. Klee, J. Cochran, M. Merrick, L.M. Blumberg, Evaluation of conditions of comprehensive two-dimensional gas chromatography that yield a near-theoretical maximum in peak capacity gain, J. Chromatogr. A 1383 (2015) 151–159. https://doi.org/10.1016/j.chroma.2015.01.031.

[20] A.L. Lee, K.D. Bartle, A.C. Lewis, A model of peak amplitude enhancement in orthogonal two-dimensional gas chromatography, Anal. Chem. 73 (2001) 1330–1335. https://doi.org/10.1021/ac001120s.

[21] C. Cordero, J. Kiefl, P. Schieberle, S.E. Reichenbach, C. Bicchi, Comprehensive two-dimensional gas chromatography and food sensory properties: Potential and challenges, Anal. Bioanal. Chem. 407 (2015) 169–191. https://doi.org/10.1007/s00216-014-8248-z.

[22] F. Stilo, C. Bicchi, A. Robbat, S.E. Reichenbach, C. Cordero, Untargeted approaches in food-omics: The potential of comprehensive two-dimensional gas chromatography/mass spectrometry, TrAC - Trends Anal. Chem. 135 (2021) 116162. https://doi.org/10.1016/j.trac.2020.116162.

[23] D. Ryan, R. Shellie, P. Tranchida, A. Casilli, L. Mondello, P. Marriott, Analysis of roasted coffee bean volatiles by using comprehensive two-dimensional gas chromatography-time-of-flight mass spectrometry, J. Chromatogr. A 1054 (2004) 57–65. https://doi.org/10.1016/j.chroma.2004.08.057.

[24] S.T. Chin, G.T. Eyres, P.J. Marriott, Identification of potent odourants in wine and brewed coffee using gas chromatography-olfactometry and comprehensive two-dimensional gas chromatography, J. Chromatogr. A 1218 (2011) 7487–7498. https://doi.org/10.1016/j.chroma.2011.06.039.

[25] F.J.M. Novaes, A.I. da Silva Junior, C. Kulsing, Y. Nolvachai, H.R. Bizzo, F.R. de Aquino Neto, C.M. Rezende, P.J. Marriott, New approaches to monitor semi-volatile organic compounds released during coffee roasting using flow-through/active sampling and comprehensive two-dimensional gas chromatography, Food Res. Int. 119 (2019) 349–358. https://doi.org/10.1016/j.foodres.2019.02.009.

[26] G.R. Lopes, S. Petronilho, A.S. Ferreira, M. Pinto, C.P. Passos, E. Coelho, C. Rodrigues, C. Figueira, S.M. Rocha, M.A. Coimbra, Insights on Single-Dose Espresso Coffee Capsules' Volatile Profile: From Ground Powder Volatiles to Prediction of Espresso Brew Aroma Properties, Foods 10 (2021) 2508. https://doi.org/10.3390/foods10102508.

[27] Y. Zou, M. Gaida, F.A. Franchina, P.H. Stefanuto, J.F. Focant, Distinguishing between Decaffeinated and Regular Coffee by HS-SPME-GC×GC-TOFMS, Chemometrics, and Machine Learning, Molecules 27 (2022). https://doi.org/10.3390/molecules27061806.

[28] A. Pua, Y. Huang, R.M. Vivian Goh, K.-H. Ee, L. Li, M. Cornuz, B. Lassabliere, L. Jublot, S.Q. Liu, B. Yu, Multidimensional Gas Chromatography of Organosulfur Compounds in Coffee and Structure–Odor Analysis of 2-Methyltetrahydrothiophen-3-one, J. Agric. Food Chem. 71 (2023) 4337–4345. https://doi.org/10.1021/acs.jafc.2c08842.

[29] E.M. Humston, J.D. Knowles, A. McShea, R.E. Synovec, Quantitative assessment of moisture damage for cacao bean quality using two-dimensional gas chromatography combined with time-of-flight mass spectrometry and chemometrics, J. Chromatogr. A 1217 (2010) 1963–1970. https://doi.org/10.1016/j.chroma.2010.01.069.

[30] P.H. Stefanuto, K.A. Perrault, L.M. Dubois, B. L'Homme, C. Allen, C. Loughnane, N. Ochiai, J.F. Focant, Advanced method optimization for volatile aroma profiling of beer

using two-dimensional gas chromatography time-of-flight mass spectrometry, J. Chromatogr. A 1507 (2017) 45–52. https://doi.org/10.1016/j.chroma.2017.05.064.

[31] M. Cialiè Rosso, E. Liberto, N. Spigolon, M. Fontana, M. Somenzi, C. Bicchi, C. Cordero, Evolution of potent odorants within the volatile metabolome of high-quality hazelnuts (Corylus avellana L.): evaluation by comprehensive two-dimensional gas chromatography coupled with mass spectrometry, Anal. Bioanal. Chem. 410 (2018) 3491–3506. https://doi.org/10.1007/s00216-017-0832-6.

[32] J. Crucello, L.F.O. Miron, V.H.C. Ferreira, H. Nan, M.O.M. Marques, P.S. Ritschel, M.C. Zanus, J.L. Anderson, R.J. Poppi, L.W. Hantao, Characterization of the aroma profile of novel Brazilian wines by solid-phase microextraction using polymeric ionic liquid sorbent coatings, Anal. Bioanal. Chem. 410 (2018) 4749–4762. https://doi.org/10.1007/s00216-018-1134-3.

[33] P.E. Sudol, M. Galletta, P.Q. Tranchida, M. Zoccali, L. Mondello, R.E. Synovec, Untargeted profiling and differentiation of geographical variants of wine samples using headspace solid-phase microextraction flow-modulated comprehensive two-dimensional gas chromatography with the support of tile-based Fisher ratio analysis, J. Chromatogr. A 1662 (2022) 462735. https://doi.org/10.1016/j.chroma.2021.462735.

[34] K.J. Johnson, R.E. Synovec, Pattern recognition of jet fuels: Comprehensive GC × GC with ANOVA-based feature selection and principal component analysis, Chemom. Intell. Lab. Syst. 60 (2002) 225–237. https://doi.org/10.1016/S0169-7439(01)00198-8.

[35] H.D. Bean, J.E. Hill, J.M.D. Dimandja, Improving the quality of biomarker candidates in untargeted metabolomics via peak table-based alignment of comprehensive two-dimensional gas chromatography-mass spectrometry data, J. Chromatogr. A 1394 (2015) 111–117. https://doi.org/10.1016/j.chroma.2015.03.001.

[36] L.C. Marney, W.C. Siegler, B.A. Parsons, J.C. Hoggard, B.W. Wright, R.E. Synovec, Tile-based Fisher-ratio software for improved feature selection analysis of comprehensive two-dimensional gas chromatography-time-of-flight mass spectrometry data, Talanta 115 (2013) 887–895. https://doi.org/10.1016/j.talanta.2013.06.038.

[37] B.A. Parsons, L.C. Marney, W.C. Siegler, J.C. Hoggard, B.W. Wright, R.E. Synovec, Tile-Based Fisher Ratio Analysis of Comprehensive Two-Dimensional Gas Chromatography Time-of-Flight Mass Spectrometry (GC × GC–TOFMS) Data Using a Null Distribution Approach, Anal. Chem. 87 (2015) 3812–3819. https://doi.org/10.1021/ac504472s.

[38] S.E. Prebihalo, G.S. Ochoa, K.L. Berrier, K.J. Skogerboe, K.L. Cameron, J.R. Trump, S.J. Svoboda, J.K. Wickiser, R.E. Synovec, Control-Normalized Fisher Ratio Analysis of Comprehensive Two-Dimensional Gas Chromatography Time-of-Flight Mass Spectrometry Data for Enhanced Biomarker Discovery in a Metabolomic Study of Orthopedic Knee-Ligament Injury, Anal. Chem. (2020). https://doi.org/10.1021/acs.analchem.0c03456.

[39] C.N. Cain, T.J. Trinklein, G.S. Ochoa, R.E. Synovec, Tile-Based Pairwise Analysis of GC × GC-TOFMS Data to Facilitate Analyte Discovery and Mass Spectrum Purification, Anal. Chem. 94 (2022) 5658–5666. https://doi.org/10.1021/acs.analchem.2c00223.

[40] S. Schöneich, G.S. Ochoa, C.M. Monzón, R.E. Synovec, Minimum variance optimized Fisher ratio analysis of comprehensive two-dimensional gas chromatography / mass spectrometry data: Study of the pacu fish metabolome, J. Chromatogr. A 1667 (2022) 462868. https://doi.org/10.1016/j.chroma.2022.462868.

[41] P.E. Sudol, G.S. Ochoa, C.N. Cain, R.E. Synovec, Tile-based variance rank initiated-unsupervised sample indexing for comprehensive two-dimensional gas chromatography-time-of-flight mass spectrometry, Anal. Chim. Acta 1209 (2022) 339847. https://doi.org/10.1016/j.aca.2022.339847.

[42] C. Thoumsin, Data Collection Methodology and Accurate Instance Rate Determination in Coffees with Potato Taste Defect (PTD), Count. Cult. Coffee. (2019) 1–9. https://counterculturecoffee.com/wp-content/uploads/2020/04/CCC-PTD-Paper-Final.pdf.

[43] Specialty Coffee Association of America, Cupping Specialty Coffee, (2015) 1–10. http://www.scaa.org/PDF/resources/ cupping-protocols.pdf (Accessed (accessed September 17, 2020).

[44] N. Caporaso, M.B. Whitworth, C. Cui, I.D. Fisk, Variability of single bean coffee volatile compounds of Arabica and robusta roasted coffees analysed by SPME-GC-MS, Food Res. Int. 108 (2018) 628–640. https://doi.org/10.1016/j.foodres.2018.03.077.

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

[46] S.C. Rutan, A. de Juan, R. Tauler, Introduction to Multivariate Curve Resolution, in: S.D. Brown, R. Tauler, B. Walczak (Eds.), Compr. Chemom., Vol. 2, Elsevier, 2009: pp. 249–259.

[47] G.S. Ochoa, S.E. Prebihalo, B.C. Reaser, L.C. Marney, R.E. Synovec, Statistical inference of mass channel purity from Fisher ratio analysis using comprehensive two-dimensional gas chromatography with time of flight mass spectrometry data, J. Chromatogr. A 1627 (2020) 461401. https://doi.org/10.1016/j.chroma.2020.461401.

[48] C.N. Cain, S. Schöneich, R.E. Synovec, Development of an Enhanced Total Ion Current Chromatogram Algorithm to Improve Untargeted Peak Detection, Anal. Chem. 92 (2020) 11365–11373. https://doi.org/10.1021/acs.analchem.0c02136.

[49] G.S. Ochoa, M.C. Billingsley, R.E. Synovec, Using solid-phase extraction to facilitate a focused tile-based Fisher ratio analysis of comprehensive two-dimensional gas chromatography time-of-flight mass spectrometry data: comparative analysis of aerospace fuel composition, Anal. Bioanal. Chem. (2022). https://doi.org/10.1007/s00216-022-04348-1.

[50] The Good Scents Company, The Good Scents Company Information System, (2018). http://www.thegoodscentscompany.com/.

[51] T.J. Trinklein, R.E. Synovec, Simulating comprehensive two-dimensional gas chromatography mass spectrometry data with realistic run-to-run shifting to evaluate the robustness of tile-based Fisher ratio analysis, J. Chromatogr. A 1677 (2022) 463321. https://doi.org/10.1016/j.chroma.2022.463321.

[52] B.C. Reaser, B.W. Wright, R.E. Synovec, Using Receiver Operating Characteristic Curves To Optimize Discovery-Based Software with Comprehensive Two-Dimensional Gas Chromatography with Time-of-Flight Mass Spectrometry, Anal. Chem. 89 (2017) 3606–3612. https://doi.org/10.1021/acs.analchem.6b04991.

[53] R. Rousseau, B. Govaerts, M. Verleysen, B. Boulanger, Comparison of some chemometric tools for metabonomics biomarker identification, Chemom. Intell. Lab. Syst. 91 (2008) 54–66. https://doi.org/10.1016/j.chemolab.2007.06.008.

[54] C.D. Brown, H.T. Davis, Receiver operating characteristics curves and related decision measures: A tutorial, Chemom. Intell. Lab. Syst. 80 (2006) 24–38. https://doi.org/10.1016/j.chemolab.2005.05.004.

[55] D.W. Hosmer, S. Lemeshow, R.X. Sturdivant, Assessing the Fit of the Model, in: Appl. Logist. Regres., 3rd ed., Wiley, Hoboken, NJ, 2013: pp. 153–226.

[56] D. Bressanello, E. Liberto, C. Cordero, B. Sgorbini, P. Rubiolo, G. Pellegrino, M.R. Ruosi, C. Bicchi, Chemometric Modeling of Coffee Sensory Notes through Their Chemical Signatures: Potential and Limits in Defining an Analytical Tool for Quality Control, J. Agric. Food Chem. 66 (2018) 7096–7109. https://doi.org/10.1021/acs.jafc.8b01340.

[57] L.F. Huang, M.J. Wu, K.J. Zhong, X.J. Sun, Y.Z. Liang, Y.H. Dai, K.L. Huang, F.Q. Guo, Fingerprint developing of coffee flavor by gas chromatography-mass spectrometry and combined chemometrics methods, Anal. Chim. Acta 588 (2007) 216–223. https://doi.org/10.1016/j.aca.2007.02.013.

[58] S. Blumberg, O. Frank, T. Hofmann, Quantitative studies on the influence of the bean roasting parameters and hot water percolation on the concentrations of bitter compounds in coffee brew, J. Agric. Food Chem. 58 (2010) 3720–3728. https://doi.org/10.1021/jf9044606.

[59] L. Poisson, I. Blank, A. Dunkel, T. Hofmann, The Chemistry of Roasting—Decoding Flavor Formation, in: Cr. Sci. Coffee, Elsevier, London, UK, 2017: pp. 273–309. https://doi.org/10.1016/B978-0-12-803520-7.00012-8.

[60] G. Strocchi, E. Bagnulo, M.R. Ruosi, G. Ravaioli, F. Trapani, C. Bicchi, G. Pellegrino, E. Liberto, Potential Aroma Chemical Fingerprint of Oxidised Coffee Note by HS-SPME-GC-MS and Machine Learning, Foods 11 (2022) 1–14. https://doi.org/10.3390/foods11244083.

[61] P.E. Sudol, G.S. Ochoa, R.E. Synovec, Investigation of the limit of discovery using tile-based Fisher ratio analysis with comprehensive two-dimensional gas chromatography time-of-flight mass spectrometry, J. Chromatogr. A 1644 (2021). https://doi.org/10.1016/j.chroma.2021.462092.

[62] V.E. de Almeida, D.D. de Sousa Fernandes, P.H.G.D. Diniz, A. de Araújo Gomes, G. Véras, R.K.H. Galvão, M.C.U. Araujo, Scores selection via Fisher's discriminant power in PCA-LDA to improve the classification of food data, Food Chem. 363 (2021). https://doi.org/10.1016/j.foodchem.2021.130296.

[63] A.C. de Oliveira Junqueira, G. V. de Melo Pereira, J.D. Coral Medina, M.C.R. Alvear, R. Rosero, D.P. de Carvalho Neto, H.G. Enríquez, C.R. Soccol, First description of bacterial

and fungal communities in Colombian coffee beans fermentation analysed using Illumina-based amplicon sequencing, Sci. Rep. 9 (2019) 1–10. https://doi.org/10.1038/s41598-019-45002-8.

[64] X. Shen, B. Wang, C. Zi, L. Huang, Q. Wang, C. Zhou, W. Wen, K. Liu, W. Yuan, X. Li, Interaction and Metabolic Function of Microbiota during the Washed Processing of Coffea arabica, Molecules 28 (2023). https://doi.org/10.3390/molecules28166092.

[65] A. Farah, M.C. Monteiro, V. Calado, A.S. Franca, L.C. Trugo, Correlation between cup quality and chemical attributes of Brazilian coffee, Food Chem. 98 (2006) 373–380. https://doi.org/10.1016/j.foodchem.2005.07.032.

[66] F. Zhao, P. Wang, R.D. Lucardi, Z. Su, S. Li, Natural sources and bioactivities of 2,4-di-tert-butylphenol and its analogs, Toxins 12 (2020) 1–26. https://doi.org/10.3390/toxins12010035.

[67] S. Fang, S. Liu, J. Song, Q. Huang, Z. Xiang, Recognition of pathogens in food matrixes based on the untargeted in vivo microbial metabolite profiling via a novel SPME/GC × GC-QTOFMS approach, Food Res. Int. 142 (2021). https://doi.org/10.1016/j.foodres.2021.110213.

Chapter 4: Detailed Chemical Compositional Analysis of a Thermally Stressed Rocket Fuel using GC×GC-TOFMS and Chemometric Data Analysis

4.1. Introduction

The performance and reliability of rocket engines is intricately linked to understanding the physiochemical properties of a given kerosene-based fuel formulation. As a result, both the chemical composition (e.g., presence and abundance of different compound classes) and thermophysical behavior (e.g., density, viscosity, heat of combustion) of a given rocket fuel is extensively studied. Typically, these physiochemical measurements are carefully made to ensure that the fuel formulation does not decompose and/or change from its as-received condition [1–5]. However, there is also a need to understand the physiochemical properties of a kerosene-based rocket fuel after exposure to the high temperatures experienced in an engine. In rocket engines, the rocket fuel is used as a heat sink to cool the nozzle before being burned in the combustion chamber [6]. The heat absorbed by the fuel from the cooling process can cause the decomposition of hydrocarbons into carbonaceous deposits ("coking"), which can result in engine failure [7–9]. Furthermore, this decomposition process can alter the physiochemical properties of the rocket fuel, such that performance will deviate from the expectations made using the fuel in its as-received condition [10–13]. For example, the presence of contaminants in kerosene-based rocket fuels (i.e., sulfur, olefins, oxygenates, and aromatics) have been previously linked to thermal instability [14–17]. However, the influence of thermal stressing temperature on the entire chemical composition of a rocket fuel has yet to be fully characterized.

Current standardized methods for characterizing fuel composition rely on the use of one-dimensional gas chromatography (1D-GC) [2,18–21], but achieving a detailed compositional

analysis of a kerosene-based fuel in an appropriate amount of time can be a challenge for 1D-GC [21]. Fortunately, comprehensive two-dimensional (2D) gas chromatography (GC×GC) is a powerful analytical method for improving the resolution of volatile and semi-volatile mixtures for compositional analysis [22]. In GC×GC, two separations with complementary selectivity are connected via a modulator, which injects effluent from the first dimension (^{1}D) column onto the second dimension (^{2}D) column. In turn, the use of two separation dimensions in GC×GC provides an increase in peak capacity compared to 1D-GC [23]. Additionally, the modulation process also increases sensitivity, which can improve the detection of low concentration species in a sample [24,25]. Coupling GC×GC with time-of-flight mass spectrometry (GC×GC-TOFMS) produces an instrumental platform with high temporal and spectral resolution for characterizing various matrices, including kerosene-based fuels. Indeed, GC×GC-TOFMS has demonstrated its ability to resolve and identify chemical species in various kerosene-based fuels [5,26–38]. In particular, the use of a polar ^{1}D column and non-polar ^{2}D column provides excellent selectivity and resolution for petroleum-based samples [39]. Thus, GC×GC-TOFMS can provide a wealth of chemical information regarding fuel composition.

The use of non-targeted chemometric methods for GC×GC-TOFMS data analysis can provide deep insights into the relationship between fuel composition and performance. Non-targeted chemometrics refers to a suite of computational algorithms designed to build predictive regression models or discover differences between samples based on the chemical information stored in instrumentally obtained data [40]. While several studies have explored the use of regression models to predict fuel performance based on GC×GC-TOFMS data [5,27,28,32,41], few have utilized chemometrics to discover the specific compounds responsible for the observed physicochemical behavior of a fuel. These studies have primarily focused on the discovery of

impurities responsible for coke deposition (e.g., polar compounds and olefins) in fuel before exposure to thermal stress [30,38,42]. To compare the collected fuel samples and discover these chemical species, we have employed chemometric methods known as tile-based Fisher ratio (F-ratio) analysis and principal component analysis (PCA) to provide a detailed chemical perspective of the thermal stressing process.

Tile-based F-ratio analysis is a supervised chemometric method that can enable the discovery of compositional differences between fuel samples [43,44]. This algorithm discovers analytes that differentiate samples based on the magnitude of their F-ratio, which is mathematically defined as the ratio of the between-class variance to the within-class variance. This analysis is performed on a per-mass channel (*m/z*) basis using small, rectangular sections (i.e., tiles) of the GC×GC-TOFMS chromatograms [43,44]. The output of tile-based F-ratio analysis is a "hit list", which ranks analytes in descending order of their F-ratio. Hence, analytes with a high F-ratio are more likely to discriminate different fuel samples from one another. Meanwhile, PCA is an unsupervised chemometric method commonly used to visualize the similarities and/or differences between samples [45]. The goal for PCA is to develop a model that defines the underlying variation within a data set. The outputs from this model are the scores, which describe the similarity between samples, and the loadings, which highlight the most influential analytes in the model (i.e., the analytes that best represent the variation in the data set) [45].

Using these analytical and computational methods, our current study aims to establish the relationship between thermal stressing temperature and the chemical composition of a kerosene-based rocket fuel. A highly paraffinic rocket fuel was exposed to four different thermal stressing temperatures (300, 500, 700, and 900 ^{o}F) using the Compact Rapid Assessment of Fuel Thermal

Integrity (CRAFTI) apparatus [46-48]. This system provides a laboratory scale experiment of fuel thermal stability at conditions relevant to rocket regenerative cooling systems [46–48]. Briefly, fuel flows through a copper test article, a section which is resistively heated (the heated zone) to temperatures that promotes the onset of fuel chemical decomposition. Thermal (and catalytic) conditions at the cooling channel inner surface accelerate reactivity with fuel constituents, resulting in the formation of carbonaceous deposits in the heated and unheated downstream surfaces of the test article. The chemical composition of these fuels after exposure to thermal stress was analyzed with GC×GC-TOFMS and tile-based F-ratio analysis. Analytes with a statistically significant concentration difference based on the thermal stressing temperature were then identified and quantified. PCA was also coupled to the tile-based F-ratio results to highlight the differences in fuel composition after exposure to various levels of heating in the CRAFTI apparatus. We envisage that the use of the tools will not only provide additional insight into the fuel decomposition occurring during thermal stress but will also provide a platform for the development and testing of new fuel formulations.

4.2. Methods and Materials

4.2.1. Fuel samples

All thermal stressing experiments were performed on a highly paraffinic rocket fuel formulations by the Air Force Research Laboratory (AFRL, Edwards AFB, CA) using the CRAFTI apparatus [46-48]. The fuels were exposed to four extreme temperatures to induce thermal stress: 300, 500, 700, and 900 ºF (Figure 4.1). The thermally stressed fuels, along with their neat counterpart, were then forwarded for analytical characterization with GC×GC-TOFMS and chemometric data analysis.

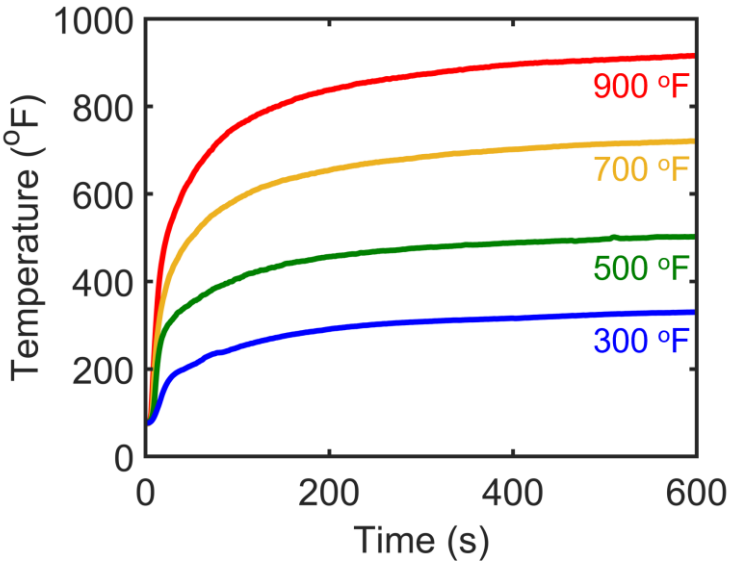


Figure 4.1. The four fuel temperatures measured at the outlet of the CRAFTI apparatus. Each line correlates to a different CRAFTI run: 300 °F – blue, 500 °F – green, 700 °F – yellow, and 900 °F – red.

4.2.2. GC×GC-TOFMS characterization

The chemical composition of thermally stressed fuel samples was analyzed using a GC×GC-TOFMS. The LECO Pegasus BT 4D GC×GC-TOFMS was equipped with an Agilent 7890B GC (Agilent Technologies, Palo Alto, CA, USA), a thermal modulator, and an L-PAL3 GC Autosampler (LECO, St. Joseph, MI, USA). A 0.5 µL injection of sample was introduced into the inlet at a 1:100 split ratio, with the inlet temperature of 275 °C. The samples were separated using a mid-polar ^{1}D column (Rxi-17Sil MS; 28 m × 0.25 mm × 0.25 µm), and a non-polar ^{2}D column (Rxi-1ms; 2 m × 0.18 mm × 0.18 µm; Restek, Bellefonte, PA, USA). Ultra-high purity Helium (Grade 5, 99.999%, Praxair, Seattle, WA, USA) was the carrier gas for the separations and was delivered at a constant flow rate of 2 mL/min. The primary oven was held at 40 °C for 1.5 min before being ramped to 200 °C at 5 °C/min, where it was held for 1 min at 200 °C. The secondary oven and thermal modulator tracked the primary oven temperature with a +20 °C and +18 °C offset, respectively. The modulation period (P_M) for the GC×GC separations was

3 s, with hot and cold pulses of 0.75 s for both stages of the thermal modulator. The transfer line to the TOFMS was held at 285 °C throughout each separation run, and the ion source was held at 225 °C. Mass spectra from *m/z* 40-334 were collected by the TOFMS at 100 Hz with an electron ionization energy of 70 eV after a 120 s acquisition delay. Three replicates of each fuel were collected.

4.2.3. Data analysis

The GC×GC-TOFMS chromatographic data was imported into Matlab 2019b (Mathworks, Inc., Natick, MA, USA) equipped with PLS Toolbox 8.9 (Eigenvector Research, Manson, WA, USA) for data analysis. The chromatograms were baseline corrected and normalized to the sum of the total ion current (TIC) chromatogram of each sample run. Tile-based F-ratio analysis was performed by comparing the fuels thermally stressed at 900 °F to those that were thermally stressed at 300 °F. A tile size of 18 s × 300 ms (^{1}D × ^{2}D) and cluster window size of 12 s × 180 ms were chosen to fully encompass the average peak widths in both dimensions. A *S/N* threshold of 10 was employed to exclude tiles with low signals from the analysis. After removing redundant hits, a final hit list was achieved by ranking the peaks in descending order of their F-ratios, which was measured at the top F-ratio *m/z*.

Class-distinguishing analytes in the hit list (true positives) were identified by calculating a *p*-value from a *t*-test (assuming unequal variances) using the tile signal at the top F-ratio *m/z*. A *p*-value < 0.05 was selected to discover these class-distinguishing analytes in the hit list. Tentative compound identifications for these true positives were made by matching the experimental mass spectrum to the NIST 11 library (National Institute of Standards and Technology, Gaithersburg, MD, USA), where a match value (MV) ≥ 800 was required for identification [49]. For situations where a given hit was unresolved from interferent compounds,

parallel factor analysis (PARAFAC) was performed to purify the mass spectrum for library matching. Accurate concentration ratios between the two thermally stressed fuel classes, [900 ºF]/[300 ºF], were calculated using pure *m/z* for each class-distinguishing hit. These pure analyte *m/z* were discovered using the signal ratio (S-ratio) algorithm [50]. This algorithm discovers pure analyte *m/z* as having a *p*-value < 0.01 and a lack-of-fit (*LOF*), a peak shape consistency metric, ≤ 20 %. To visualize the locations of the class-distinguishing analytes discovered by F-ratio analysis, a "stitch" chromatogram was constructed [38]. The stitch chromatogram was created by extracting an 18 s × 300 ms tile centered upon each analyte location at its pure analyte *m/z*, removing any background noise within the tile, and inserting those tiles back into an empty chromatogram at their original retention time location. Using these pure analyte *m/z*, peak areas were also obtained for the true positive hits in all samples (300, 500, 700, 900 °F). PCA was performed using the quantified class-distinguishing hits to describe how these different thermal stressing conditions affect fuel composition. To quantify the separation between the sample classes on the PCA scores plot, a degree-of-class separation (DCS) was calculated [51]. The DCS metric is defined as

$$DCS = \frac{D_{A,B}}{\sqrt{s_A^2 + s_B^2}} \tag{4.1}$$

where $D_{A,B}$ is the Euclidean distance between the centroids of two sample classes, A and B, and s^2 is the variance observed in the distances of each score from the centroid of a given sample class [51].

4.3. Results and Discussion

Figure 4.2A-C shows the TIC chromatograms of the original rocket fuel (A), after its exposure to temperatures of 300 ºF (B), and 900 ºF (C) using the CRAFTI apparatus. Use of a reverse column configuration (polar ^{1}D × non-polar ^{2}D) for the GC×GC-TOFMS data collection

provides excellent separation of the main compound classes (paraffins, cyclic paraffins, and aromatics) in a kerosene-based fuel. Visually, the chromatograms of the original fuel and after exposure to 300 ºF appear similar in signal while the fuel exposed to 900 ºF contains many additional analytes in the chromatographic region between 0 – 10 min on the ^{1}D and 0 – 2 s on the ^{2}D. This region of the chromatograms is primarily home to paraffins and olefins with low molecular weights. Figure 4.2D-F provides a zoom-in on this area in the chromatograms (original – D, 300 ºF – E, and 900 ºF – F) with the color scale adjusted to enhance the signal of these peaks. Here, it is observed that there are chromatographic differences between all three of these fuels, with the overall signal of this area increasing as exposure to thermal stress increases. In principle, every peak in these chromatograms could be identified and quantified to create a chemical profile of these fuels. However, to efficiently discover these compositional differences, a chemometrics-based approach for chemical analysis was employed.

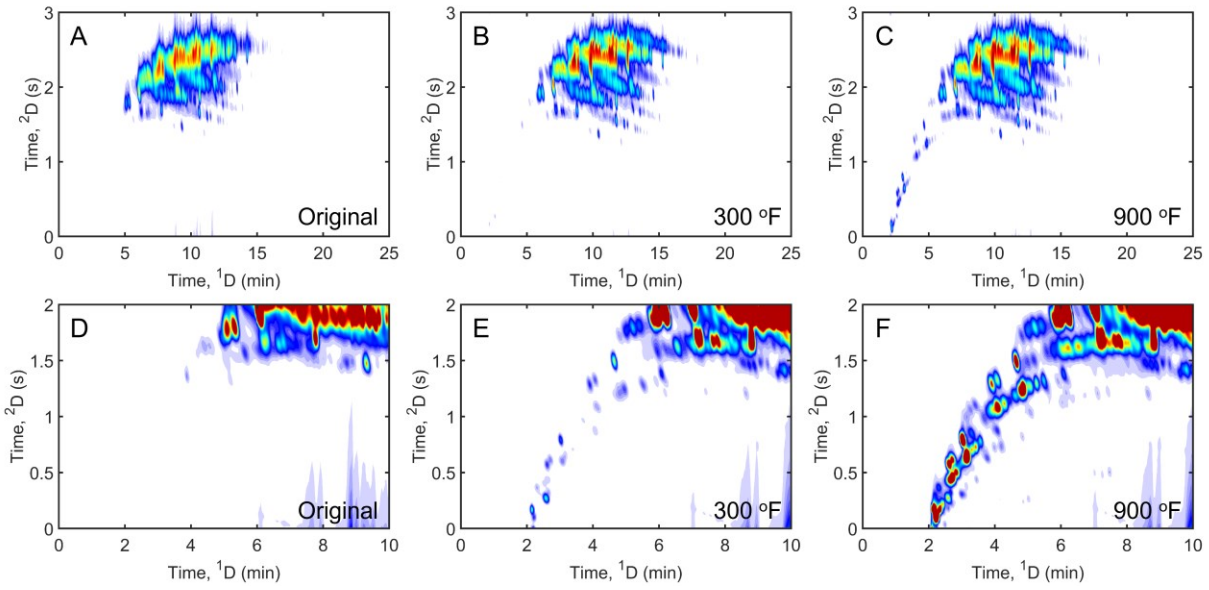


Figure 4.2. (Top row) GC×GC-TOFMS TIC chromatograms of the (A) original fuel and fuel after exposure to temperatures of (B) 300 ºF and (C) 900 ºF. (Bottom row) Use of a different color scale and zoom in on the chromatographic region between 0 – 10 min of ^{1}D and 0 – 2 s on the ^{2}D to highlight the compositional differences between the (D) original fuel and fuels stressed at (E) 300 ºF and (F) 900 ºF.

Tile-based F-ratio analysis was first performed by comparing the fuels exposed to temperatures of 300 ^{o}F and 900 ^{o}F to relate differences in composition to degree of thermal stress exposure. The resulting F-ratio distribution for this comparison is shown in Figure 4.3A, with the vertical dashed line representing a log(F-ratio) threshold of 1.75 or an F-ratio threshold of 57. It was determined that peaks above this threshold had a statistically significant (p-value < 0.05) difference in concentration between the two thermal stress temperatures while peaks below this threshold had a similar concentration in both fuels. In all, 92 analyte peaks were found to have significant concentration differences between the two fuels. To illustrate the locations of these significant peaks discovered by F-ratio analysis, their retention times were plotted on top of the fuel exposed to 900 ^{o}F (Figure 4.3B). Note, the size of the retention time markers for each peak represents the magnitude of the F-ratio. Figure 4.3B demonstrates that many of the peaks with large F-ratios are located near and/or within the chromatographic region highlighted earlier in Figure 4.2D-F (0 – 10 min on the ^{1}D and 0 – 2 s on the ^{2}D). However, a portion of the peaks discovered are also located at later ^{1}D retention times, where the signal for those analytes is too low to appear in the TIC chromatogram. To improve the visualization of these F-ratio results, a stitch chromatogram [38] was constructed by extracting the peak signal from a selective *m/z* for each peak discovered and removing the superfluous peaks that were unaffected by thermal stress exposure. Figure 4.3C shows the resulting stitch chromatogram for the 92 peaks discovered by F-ratio analysis.

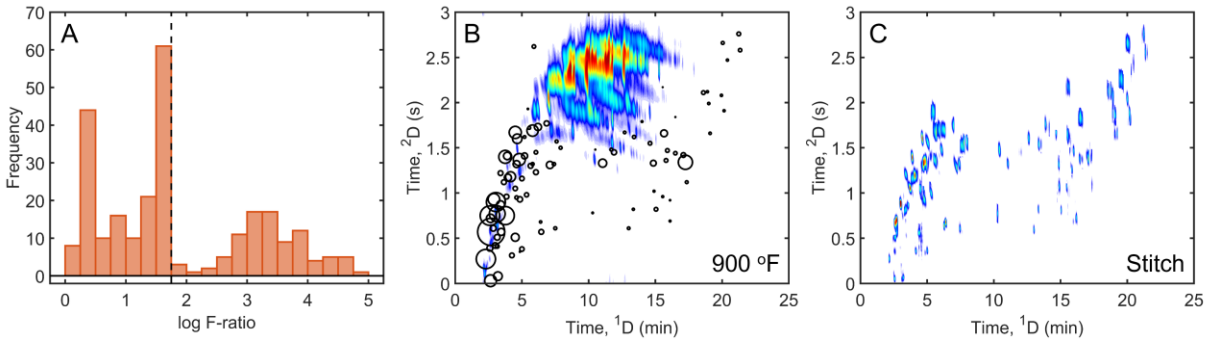


Figure 4.3. Summary of tile-based F-ratio results. (A) Distribution of F-ratios observed for the comparison of the original fuel stressed at 300 ºF and 900 ºF. Analytes with a log F-ratio greater than 1.75 (or F-ratio greater than 57) were found to be class-distinguishing. (B) GC×GC-TOFMS TIC chromatogram of the fuel exposed to 900 ºF with retention time markers highlighting the location of the 92 class-distinguishing analytes. The size of the marker indicates the magnitude of the F-ratio. (C) Stitch chromatograms of the 92 discovered analytes. For each analyte, the signal at its pure analyte *m/z* is plotted.

Table 4.1 provides the tentative mass spectrum identifications for these class-distinguishing peaks. A total of 36 olefins, 33 paraffins, 11 aromatics, and 12 oxygenated compounds were discovered. It was expected that olefins would be many of the analytes with significant concentration differences based on the heat exposure since this hydrocarbon class is one of the primary causes of thermal instability in rocket fuels and the formation of carbonaceous deposits in regenerative cooling engines [14-17]. A concentration ratio between the two thermally stressed fuel classes ([900 ºF]/[300 ºF]), was calculated using a pure *m/z* for each analyte. Those *m/z* along with the metrics used to determine their purity (*p*-value and *LOF*) are also provided in Table 4.1. These concentration ratios highlight that the formation of all 92 analytes increases as the temperature inside the CRAFTI apparatus increases. This result explains why the signals for these peaks are more prominent in the TIC chromatogram of the fuel exposed to 900 ºF (Figure 4.2F) versus the fuel exposed to 300 ºF (Figure 4.2E).

Table 4.1. The 92 analytes discovered by F-ratio analysis with a statistical difference in concentration between the fuels stressed at 300 °F and 900 °F. The hit list is ranked in descending order of F-ratio. A concentration ratio for each analyte was calculated as [900 °F]/[300 °F] using a pure *m/z* by applying the S-ratio algorithm [50]. Metrics for determining *m/z* purity (*p*-value and *LOF*) are also reported.

Hit Number	Time, ^{1}D (min)	Time, ^{2}D (s)	F-ratio	Compound	MV	S-ratio *m/z*	*p*-value	*LOF* (%)	[900 °F]/ [300 °F]
1	2.7	0.57	97354	3-Heptene	861	71	7.82E-04	2.93	11.60
2	2.6	0.75	49884	Octane	863	114	1.60E-03	4.48	16.08
3	2.3	0.27	48902	2-Heptene	956	100	8.87E-04	8.81	11.13
4	3.05	0.9	47571	Nonane	902	114	9.91E-04	16.0	18.87
5	3.75	0.75	43610	1-Methyl-2-methylenecyclohexane	901	82	1.57E-03	5.90	130.44
6	3.15	0.77	33211	1-Octene	930	70	7.82E-04	14.4	11.60
7	17.25	1.34	26609	Tricyclo[7.3.0.0(2,6)]dodecane, *trans-syn-trans-*	829	80	2.47E-02	1.38	2.27
8	3.75	1.4	22049	Heptane, 4-ethyl-	819	85	1.45E-03	3.34	6.19
9	4.5	1.67	19436	Decane	897	70	1.19E-03	16.7	3.27
10	4.8	1.37	19168	3-Nonene	913	100	1.45E-03	5.48	6.19
11	2.65	0.03	18248	Benzene	866	78	2.97E-03	12.6	3.03
12	2.9	0.93	17722	Hexane, 2,4-dimethyl-	845	85	6.22E-04	14.6	17.44
13	5.8	1.69	16843	1-Decene	864	51	2.64E-03	7.32	4.98
14	4.15	1.18	13077	1-Nonene	876	45	2.33E-03	15.3	23.97
15	4.65	1.6	11827	Heptane, 2,4,6-trimethyl-	905	70	2.46E-03	13.4	3.46
16	3.2	0.08	9864	Pentanal	879	58	1.43E-03	14.8	7.62
17	2.65	0.72	9124	Octane, 4-methyl-	882	85	1.80E-03	11.7	15.05
18	11.05	1.33	8689	4,7-Methano-1H-indene, octahydro-	831	121	2.00E-03	4.05	5.84
19	3.4	0.87	8374	2-Octene	946	112	2.22E-03	18.5	29.22
20	4.5	0.51	8227	1,3-Cyclopentadiene, 1,2-dimethyl-	836	79	2.23E-04	12.7	6.32
21	3.9	1.41	7247	Octane, 2-methyl-	891	113	2.13E-03	1.19	8.80
22	7.1	1.31	6803	1-Cyclobutanone,2-(2-methyl-1-propenyl)	832	95	1.71E-03	11.6	4.11
23	6.2	1.73	6624	1-Octene, 3,7-dimethyl-	875	69	2.75E-03	16.1	3.87
24	3.95	1.17	6579	1-Octene, 3-methyl-	893	96	1.24E-03	8.21	11.88
25	3.45	0.57	6441	Cyclohexene, 3-methyl-	868	96	2.27E-03	14.5	34.73
26	4.6	1.32	6339	2-Nonene	884	126	1.97E-03	15.6	9.12
27	15.65	1.66	6173	Cyclododecene	844	166	1.02E-03	13.8	2.16
28	5.2	1.41	4805	4-Nonene	942	98	1.83E-03	4.90	46.49
29	2.6	0.39	4670	Cyclohexane, methyl-	918	85	1.70E-03	16.4	2.49
30	2.85	0.61	4376	1-Heptene, 5-methyl-	919	85	2.31E-03	15.0	15.42

31	14.85	1.33	4315	Tetrahydrofuran, 2-methyl-5-pentyl-	802	85	1.77E-03	5.98	9.33
32	3.15	0.51	4311	1-ethylcyclopentene	823	54	2.62E-03	11.1	153.39
33	3	0.42	4074	Cyclohexane, methylene-	873	54	1.86E-03	14.8	190.80
34	4.85	0.93	3726	4-Ethylcyclohexene	824	54	2.22E-03	19.9	365.24
35	11.9	1.45	3438	Tricyclo[3.3.1.1(3,7)]decane, 1-nitro-	824	99	1.35E-03	15.2	12.77
36	6.85	1.77	3378	3-Decene	860	96	3.81E-03	11.9	2.15
37	6.05	1.23	3120	Propylidencyclohexane	864	54	2.55E-03	2.47	8.91
38	6.45	0.57	3110	*p*-Xylene	834	91	2.91E-03	12.8	2.99
39	5.65	1.3	3105	3-Heptene, 2,6-dimethyl-	817	126	9.02E-03	18.6	2.30
40	3.4	1.22	3103	Heptane, 2,5-dimethyl-	907	99	2.70E-03	3.25	5.82
41	3.65	1.12	3088	Cyclohexane, 1,2,4-trimethyl-	847	111	9.80E-03	18.3	2.56
42	3.85	0.96	2992	Cyclohexane, 1,3,5-trimethyl-	882	111	2.41E-03	15.2	2.40
43	4.25	0.31	2901	Toluene	908	92	3.04E-03	15.1	3.49
44	17.1	1.42	2873	1-Ethyladamantane	825	79	1.33E-03	7.19	2.32
45	5	1.16	2645	1-propenylcyclohexane	934	109	2.67E-03	17.4	308.79
46	5	0.38	2553	Hexanal	851	56	5.91E-03	1.36	9.74
47	4.5	1.05	2543	*cis*-1,4-Dimethyl-2-methylenecyclohexane	832	81	2.35E-03	7.87	200.14
48	5.35	0.98	2302	Cyclohexane, ethylidene-	850	81	1.01E-03	10.5	17.45
49	3.15	0.41	1912	3-propylcyclohexene	894	55	1.05E-03	1.07	2.98
50	21.25	2.76	1847	Pentadecane, 2,6,10-trimethyl-	859	113	4.50E-03	14.5	2.83
51	15.9	1.47	1824	1-Adamantanemethanol	893	135	4.19E-03	2.48	2.22
52	15.8	1.36	1806	4,7-Ethano-1H-indene, octahydro-	827	150	3.86E-03	5.50	2.17
53	18.6	2.11	1799	1,1-Bicyclohexyl, 2-(2-methylpropyl)-, *trans*-	801	139	2.37E-03	3.57	5.40
54	15.05	0.82	1751	1H-Indene, 2,3-dihydro-5-methyl-	866	115	1.89E-03	7.81	2.50
55	14.6	1.55	1732	2(1H)-Naphthalenone, octahydro-8a-methyl-, *cis*-	801	71	8.56E-03	2.44	1.99
56	6.05	1.45	1654	Cyclohexane, propyl-	889	126	2.96E-03	13.2	2.24
57	5.9	2.62	1639	*trans*-1-Butenylcyclopentane	826	124	2.54E-03	11.0	17.31
58	4.6	0.95	1534	Cyclopentene, 3-propyl-	860	81	7.83E-05	9.39	35.90
59	21.35	2.58	1530	Tetradecane, 6,9-dimethyl-	866	126	7.84E-03	13.9	2.66
60	16.5	1.44	1525	1,4-Dimethyladamantane	838	149	1.47E-03	6.08	2.23
61	7.4	1.32	1401	1H-Indene, octahydro-, *trans*-	880	124	1.82E-03	16.6	2.39
62	13.4	1.62	1359	Naphthalene, decahydro-2-methyl-	878	152	1.62E-03	14.7	2.11
63	5.8	1.32	1236	1,2-Dipropylcyclopropene	820	95	2.28E-03	11.1	18.41
64	7.5	0.61	1209	Heptanal	870	70	5.76E-03	18.3	8.06
65	7.9	1.5	1175	Cyclooctene, 1,2-dimethyl-	811	123	1.17E-03	15.6	2.65
66	20.05	2.66	1148	Nonadecane	849	85	1.14E-03	14.5	2.24
67	7.6	1.57	1135	Bicyclo[4.1.0]heptane, 3,7,7-trimethyl-, (1à,3à,6à)-	856	138	8.26E-03	12.2	2.94

68	11.7	1.48	1074	*cis*-3-Methyl-endo-tricyclo[5.2.1.0(2.6)]decane	832	81	1.83E-03	5.24	2.05
69	17.35	1.12	1041	6-Dodecanone	811	58	5.73E-04	19.7	3.84
70	5.55	1.72	1036	2-Decene	862	83	4.11E-03	10.0	2.97
71	13	0.61	917	Indene	921	115	5.19E-03	18.0	2.94
72	5.2	1.62	916	5-methyl-1-undecene	818	111	3.16E-03	3.69	11.48
73	12.7	1.67	853	Decalin, *syn*-1-methyl-, *cis*-	828	152	1.35E-03	17.9	2.09
74	19.1	1.66	745	1,1-Bicyclohexyl, 2-methyl-, *cis*-	843	180	1.01E-03	19.5	2.39
75	6.4	0.68	703	Ethylbenzene	933	91	3.38E-03	1.51	2.04
76	19	1.99	680	Cyclopentaneacetaldehyde, 2-formyl-3-methyl-à-methylene-	801	151	5.65E-03	9.15	2.41
77	19.95	2.08	654	Cyclohexane, octyl-	830	196	1.83E-03	16.8	2.57
78	20.4	2.47	629	Cyclohexane, [6-cyclopentyl-3-(3-cyclopentylpropyl)hexyl]-	801	125	6.06E-03	19.5	2.75
79	18.9	2.12	602	Cyclohexane, 1-(cyclohexylmethyl)-2-methyl-, trans-	801	97	1.38E-03	5.46	2.36
80	13.35	0.82	598	1-Phenyl-1-butene	880	117	9.31E-03	5.74	2.55
81	20.15	1.91	577	1H-Indene, octahydro-2,2,4,4,7,7-hexamethyl-, *trans*-	807	193	4.22E-03	18.4	2.45
82	15.55	0.96	515	1H-Indene, 2,3-dihydro-4-methyl-	881	160	5.65E-03	6.96	2.44
83	14.9	1.02	480	Benzene, 1,4-*bis*(1-methylethyl)-	886	158	7.95E-03	8.70	2.11
84	16.1	0.69	437	1H-Indene, 3-methyl-	861	130	8.97E-03	7.05	2.75
85	14	1.79	419	Cycloundecene, 1-methyl-	800	151	1.05E-03	18.9	2.98
86	10.3	0.78	321	Octanal	840	84	2.20E-06	6.60	7.14
87	5.45	1.93	220	Heptane, 3-ethyl-2-methyl-	882	98	9.94E-03	13.8	2.15
88	16.1	0.92	202	1H-Indene, 2,3-dihydro-1,6-dimethyl-	816	131	1.42E-03	18.8	2.49
89	4.85	0.35	153	2-Hexanone	854	58	1.35E-03	1.61	4.35
90	10.5	1.48	92	Naphthalene, decahydro-, *trans*-	914	138	1.45E-03	9.74	2.08
91	16.5	1.84	82	Cyclododecene, 1-methyl-	807	180	6.25E-03	17.6	2.21
92	15.55	2.17	58	1-Tridecene	829	182	1.20E-03	13.3	2.12

PCA was then performed to visualize the compositional differences between the fuels based on the peaks discovered by F-ratio analysis. The scores plot in Figure 4.4A demonstrates that the different fuel samples cluster together based on their thermal stress conditions (or lack thereof). In fact, when comparing the scores of the different fuel samples on PC 1, it is evident that a trend in thermal stressing temperature is captured by PCA. However, it is important to note that the replicates for the fuels stressed at 300 ºF and 500 ºF are overlapped on the scores plot. In fact, the separation between each class and their nearest neighbor on the scores plot can be

quantified using the DCS metric [51]. Table 4.2 lists the DCS between the different fuel pairs. From Table 4.2, the DCS between the fuels stressed at 300 °F and 500 °F is the lowest among the different fuel pairs. This calculation allows us to conclude that the chemical compositions of these fuels are similar. Meanwhile, for the original fuel (prior to thermal stressing) and the fuel after thermal stressing at 900 °F and 700 °F, their chemical compositions are the most different in this sample set.

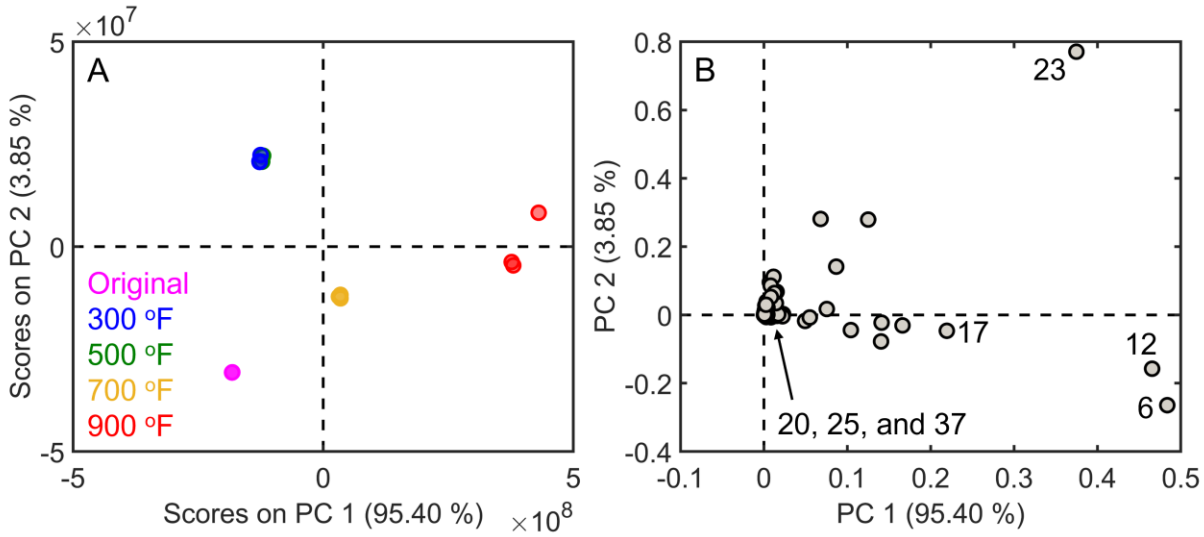


Figure 4.4. Results from PCA using the peak areas for the 92 analytes discovered by F-ratio analysis. These peak areas were measured using a pure *m/z* for each analyte. (A) Scores plot obtained from PCA, where each marker corresponds to a chromatographic replicate of the different fuel samples (original – pink, 300 °F – blue, 500 °F – green, 700 °F – yellow, and 900 °F – red). (B) Loadings plot for the model shown (A), where each gray dot corresponds to one of the analytes discovered. Three analytes centered around the origin (Hits #20, 25, and 37) and four analytes highly loaded on PC 1 (Hits #6, 12, 17, and 23) are labeled.

Table 4.2. Degree-of-class separation (DCS) calculations for the nearest neighbor fuel pairs on the PCA scores plot shown in Figure 4.4A.

Fuel Pair	DCS
900 °F vs 700 °F	34.7
700 °F vs 500 °F	208
500 °F vs 300 °F	8.0
300 °F vs Original	150

Since Figure 4.4A shows that PC 1 captures the effect of the different thermal stress conditions, the 2D loadings plot shown in Figure 4.4B can identify which analytes in the model are contributing to this sample differentiation. Note, each circle on the 2D loadings plot in Figure 4.4B represents one of the 92 class-distinguishing analytes discovered by F-ratio analysis. For instance, the analytes with positive loadings on PC 1 are more likely to have higher concentrations in the fuels exposed to temperatures of 700 ^{o}F and 900 ^{o}F since these fuels had positives scores on PC 1. Some analytes that are responsible for the sample differentiation on the scores plot includes 1-octene (Hit #6), 2,4-dimethylhexane (Hit #12), 4-methyloctane (Hit #17), and 3,7-dimethyl-1-octene (Hit #23). Conversely, analytes that clustered closely around (0,0) on the 2D loadings plot, such as 1,2-dimethyl-1,3-cyclopentadiene (Hit #20), 3-methylcyclohexene (Hit #25), and propylidencyclohexane (Hit #37), have a minimal contribution to the clustering of the fuel samples on the scores plot despite having a p-value < 0.01. This result indicates that some of the analytes discovered by F-ratio analysis also have more subtle differences in concentration between the different fuel samples. For understanding the effects of thermal stress to kerosene-based rocket propellants, it is important to explore both the analytes that are, and are not, highly loaded in the PCA model.

Figure 4.5 provides bar graphs of the average peak area of the three analytes that had minimal contributions to the PCA model in Figure 4.4: (A) 1,2-dimethyl-1,3-cyclopentadiene, (B) 3-methylcyclohexene, and (C) propylidencyclohexane. Notably, for each of these analytes, the peak area measured in the fuel after exposure to 900 ^{o}F is the largest out of all the other thermal stress conditions. Meanwhile, Figure 4.5A shows that the signal for 1,2-dimethyl-1,3-cyclopentadiene appears to be relatively constant among all the other thermal stress conditions. Similarly, Figure 4.5B-C shows that the peak area of 3-methylcyclohexene and

propylidencyclohexane is slightly higher for the fuel exposed to 700 ºF, but their signals in the remaining fuel samples are approximately the same. Based on the tentative identifications made by mass spectrum matching, all three of these analytes are olefins (Table 4.1). Previous work investigating the olefin formation due to thermal stressing of kerosene-based fuels demonstrated that olefins are primarily formed at temperatures above 350 ºC (or 662 ºF) [10,52-55]. At these high temperatures, pyrolysis is the main reaction mechanism contributing to fuel degradation, where long paraffinic chains are converted into smaller paraffins and olefins [7,56,57]. These pyrolysis products ultimately lead to the fouling of regenerative cooling engines due to the buildup of carbonaceous deposits [10,52-55]. The results shown in Figure 5 are in agreement with these previous studies because the signals for these three olefins did not start to increase until the rocket propellent was exposed to temperatures of 700 ºF and above. However, while these analytes are significant contributors to the formation of carbonaceous fouling deposits, the PCA model in Figure 4.4 indicates that these analytes, along with others centered around the origin on the loadings plot, are not the primary chemical differences between the fuel samples.

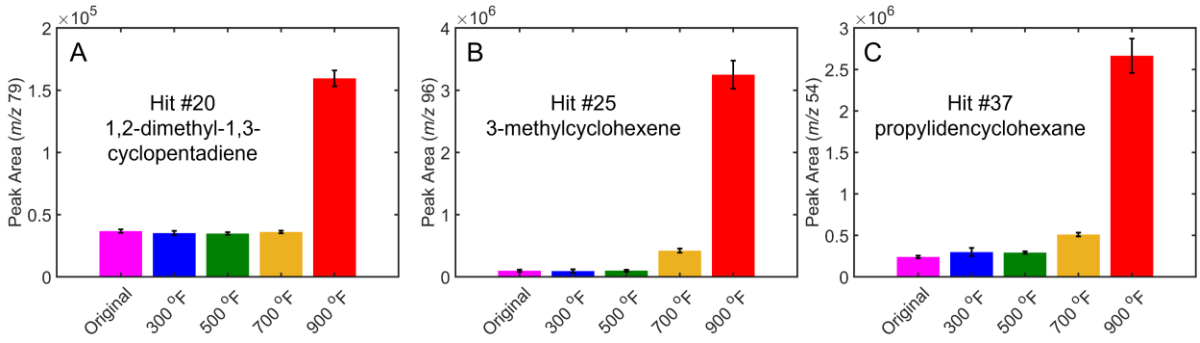


Figure 4.5. Bar graphs displaying the average peak area for (A) 1,2-dimethyl-1,3-cyclopentadiene, (B) 3-methylcyclohexene, and (C) propylidencyclohexane measured in the original fuel sample and fuels after exposure to 300, 500, 700, and 900 ºF. All peak area measurements were made using a pure analyte *m/z*. These three analytes were centered at the origin in the PCA loadings plot, provided in Figure 4.4B.

To uncover these primary chemical differences, Figure 4.6 shows the average peak area for four analytes that were highly loaded in the PCA model (Figure 4.4) based on thermal stress conditions. These four analytes correspond to: (A) 1-octene, (B) 2,4-dimethylhexane, (C) 4-methyloctane, and (D) 3,7-dimethyl-1-octene. Based on their tentative compound identification, these analytes belong to both the olefin and paraffin hydrocarbon classes (Table 4.1). Figure 6 shows that any exposure of the fuel to high temperatures increases the signal for these analytes above the signal observed in the original fuel sample. Similar to the exemplary analytes shown in Figure 4.5, the largest signals for the four analytes in Figure 4.6 are observed in the fuels thermally stressed at 700 ^{o}F and 900 ^{o}F. This result is expected since the formation of olefins and low molecular weight paraffins due to the thermal decomposition of kerosene-based fuels have been previously documented within this temperature regime [10,52-55,58,59]. However, Figure 4.6 also demonstrates that these analytes are also formed in fuels thermally stressed at 300 ^{o}F and 500 ^{o}F at approximately the same concentration. This result explains why these fuels clustered closely together on the PCA scores plot (Figure 4.4A) and had a low DCS (Table 4.2). Furthermore, the formation of these analytes suggests that the thermal decomposition of kerosene-based fuels via pyrolysis occurs at lower temperatures than previously suggested in the literature [7,56,57]. Thus, there is a need to ensure the thermal stability of fuels at lower temperatures to prevent the formation of carbonaceous deposits inside regenerative cooling engines.

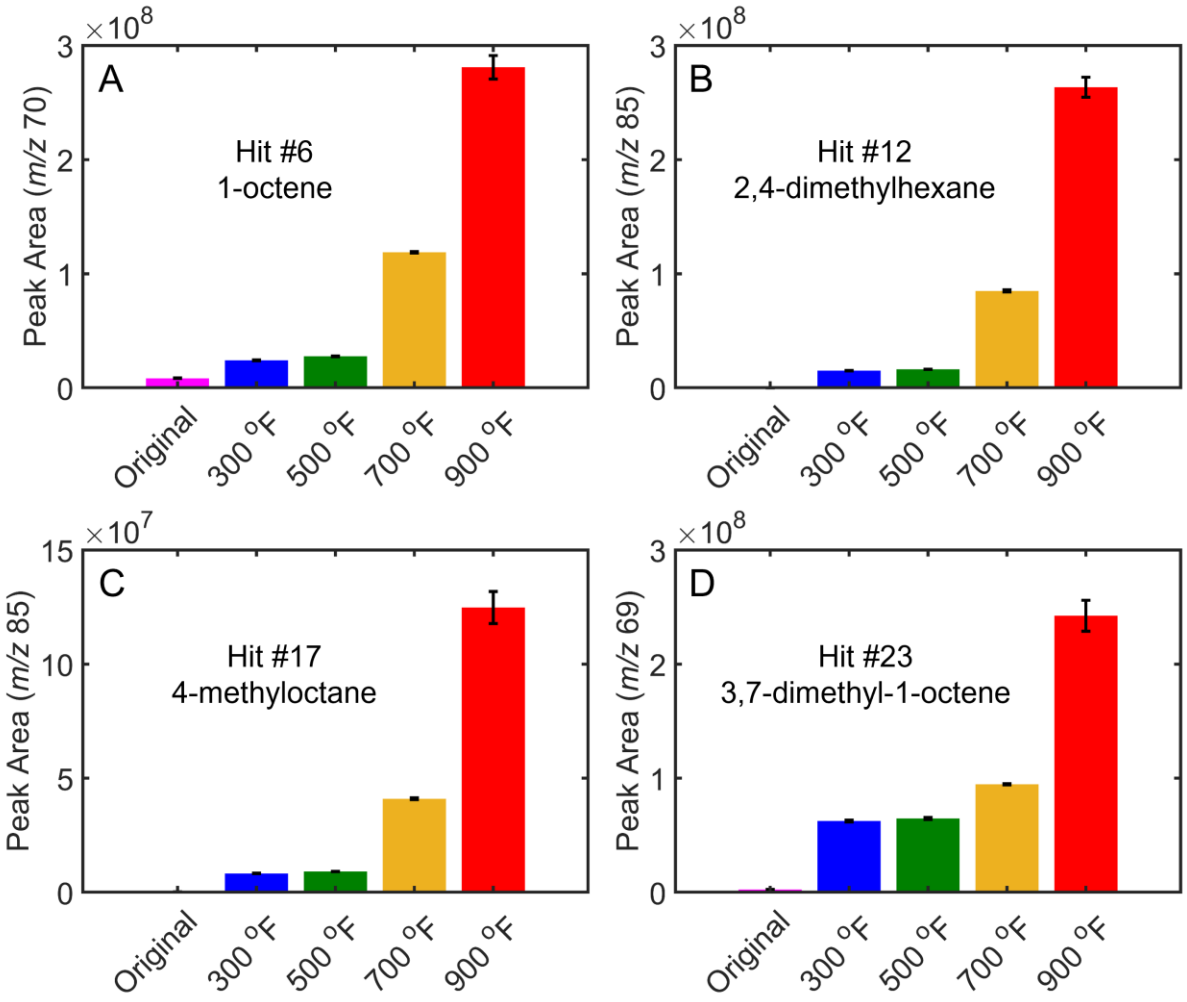


Figure 4.6. Bar graphs displaying the average peak area for (A) 1-octene, (B) 2,4-dimethylhexane, (C) 4-methyloctane, and (D) 3,7-dimethyl-1-octene measured in the original fuel sample and fuels after exposure to 300, 500, 700, and 900 °F. All peak area measurements were made using a pure analyte *m/z*. These four analytes were highly loaded on PC 1 as indicated in Figure 4.4B.

While the PCA model developed in Figure 4.4 was useful for highlighting the major chemical differences in fuel composition between the different thermal stress conditions, the model was not capable of distinguishing the fuels thermally stressed at 300 °F and 500 °F. The similar clustering of these samples on the PCA scores plot occurred because the minor chemical differences between the fuels exposed to 300 °F and 500 °F were overshadowed by the larger differences when analyzing all the fuel conditions together. Therefore, to uncover these minor

fuel differences, PCA was performed again using only the samples exposed to 300 ºF and 500 ºF (Figure 4.7A-B). The new DCS between these two samples was 8.6 (Figure 4.7A), a marginal increase compared to the old DCS value of 8.0 (Table 4.2). It was expected that the DCS for this fuel pair would not drastically change between the two PCA models since the minor replicate-to-replicate differences between the fuels are now magnified in Figure 4.7A due to the absence of all other chemical information. However, the loadings plot in Figure 4.7B is still highly informative for understanding the differences between these two fuels. Figure 4.7B shows that two analytes are highly loaded on PC 1: 2-methyl-5-pentyl-tetrahydrofuran (Hit #31) and 1-nitrotricyclo[3.3.1.1(3,7)]decane (Hit #35). Both analytes discovered by F-ratio analysis are oxygenates, which also contribute to the formation of carbonaceous deposits in regenerative cooling engines [14-17]. Figure 4.7C-D shows the bar graphs of the peak area for (C) 2-methyl-5-pentyl-tetrahydrofuran and (D) 1-nitrotricyclo[3.3.1.1(3,7)]decane measured in each of the thermal stress conditions. For both analytes, the peak area in the fuels thermally stressed at 500, 700, and 900 ºF is significantly higher than the peak areas measured in the original fuel and fuel exposed to 300 ºF. The formation of oxygenates from fuel decomposition is due to autoxidation, where the dissolved oxygen within the fuel reacts with hydrocarbon species to form heteroatomic species. Fuel autoxidation primarily occurs between fuel temperatures of 150 ºC – 350 ºC (or 302 ºF – 662 ºF) [7,56,57]. This temperature regime for fuel autoxidation correlates to the sharp increase in concentration of 2-methyl-5-pentyl-tetrahydrofuran (Figure 4.7C) and 1-nitrotricyclo[3.3.1.1(3,7)]decane (Figure 4.7D) observed in the fuels exposed to 500 ºF and 700 ºF. Furthermore, at temperatures above 350 ºC (662 ºF), oxygenated products begin to decompose, and pyrolytic reactions are the primary mechanism for fuel degradation [7,56,57]. Hence, this decomposition of the oxygenated products is a potential reason for the decrease in

signal observed in Figure 4.7C-D when comparing the fuels exposed to temperatures of 700 °F and 900 °F. The formation of these oxygenated species at temperatures greater than 500 °F can also contribute to coke deposition and ultimate rocket engine failure.

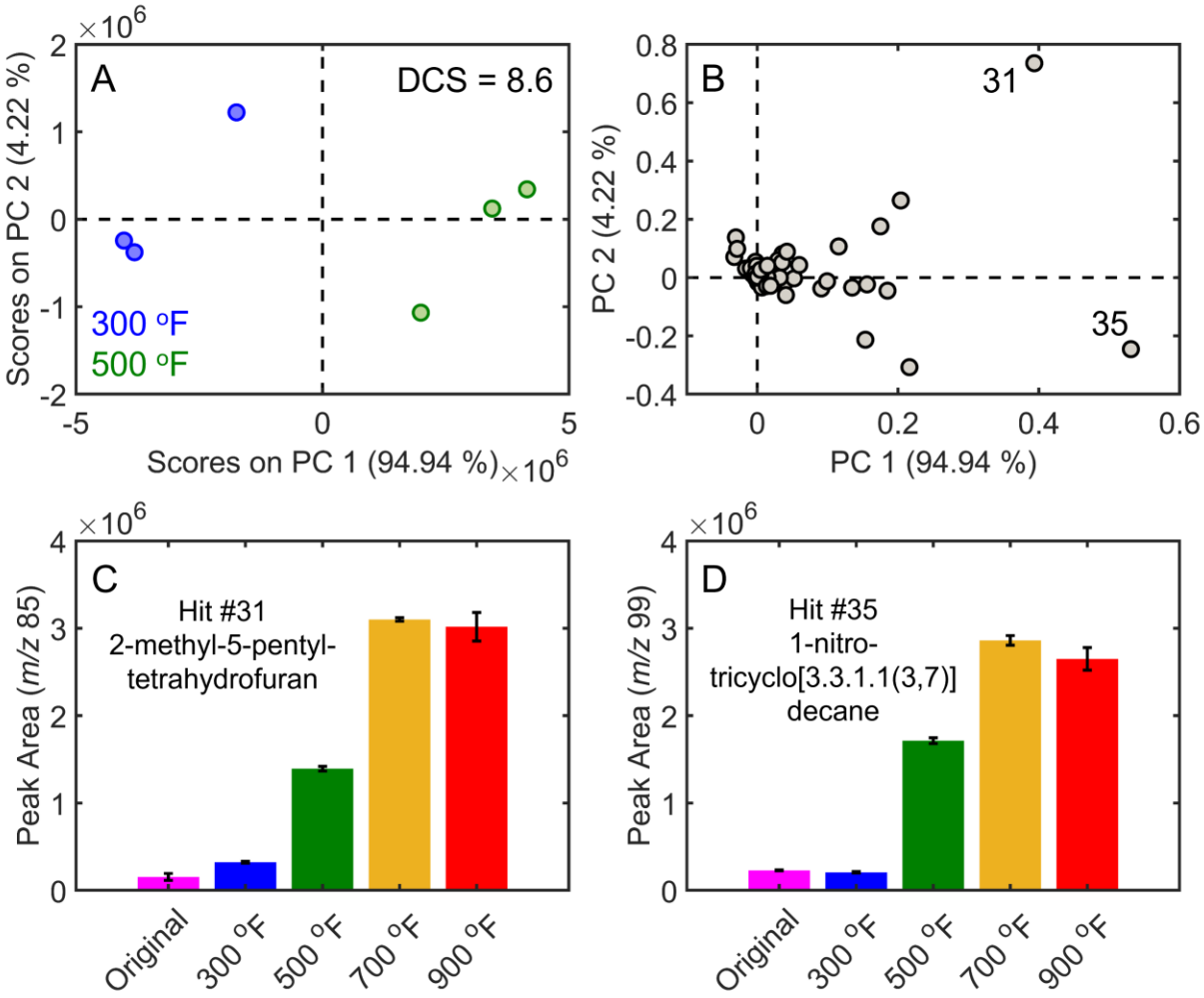


Figure 4.7. Results from PCA of the fuels thermally stressed at 300 °F (blue) and 500 °F (green). (A) Scores plot obtained from PCA of these two fuels using the peak areas for the 92 analytes discovered by F-ratio analysis. A DCS of 8.6 was calculated. (B) Loadings plot for the model shown (A), where each gray dot corresponds to one of the analytes discovered. Two analytes highly loaded on PC 1 (Hit #31 and 35) are labeled. (C) Bar graph displaying the average peak area for 2-methyl-5-pentyl-tetrahydrofuran (Hit #31) measured in the original fuel sample and fuels after exposure to 300, 500, 700, and 900 °F using a pure analyte *m/z*. (D) Bar graph displaying the average peak area for 1-nitro-tricyclo[3.3.1.1(3,7)]decane (Hit #35) measured in the original fuel sample and fuels after exposure to 300, 500, 700, and 900 °F using a pure analyte *m/z*.

4.4. Conclusion

The overarching goal of this research is to better understand and enhance fuel performance by understanding the effects of thermal stress on chemical composition. Using the CRAFTI apparatus, a highly paraffinic rocket fuel formulation was exposed to temperatures of 300, 500, 700, and 900 ºF and the chemical composition of these fuels were characterized with GC×GC-TOFMS. Tile-based F-ratio analysis was employed to discover analytes with a significant concentration difference between the fuels exposed to 300 ºF and 900 ºF. A total of 92 analytes, mainly paraffins and olefins, were discovered to have a significant concentration difference (p-value < 0.01) between these two fuels. Using the 92 analytes discovered, a PCA model of all thermal stress conditions was developed. This PCA model revealed that several fuel decomposition products were formed at thermal stress temperatures as low as 300 ºF despite current literature suggesting that these products are only formed at temperatures above 662 ºF. Furthermore, a separate PCA model comparing the fuels exposed to 300 ºF and 500 ºF revealed some of the finer differences in their chemical composition that were swamped in the original PCA model using all the fuel samples. In fact, the formation of oxygenated compounds at temperatures above 500 ºF was highlighted by this pairwise PCA model. Overall, the information highlighted in this work can be used to further tailor chemical composition of kerosene-based rocket fuels to achieve thermal stability. Future work should continue to understand how the chemical composition of different aerospace fuels are affected by different thermal stress temperatures and how temperature affects the two main reactions (autoxidation and pyrolysis) associated with fuel decomposition.

4.5. References

[1] T.J. Bruno, M.L. Huber, E.W. Lemmon, Effect of RP-1 compositional variability on thermophysical properties, Energy Fuels. 23 (2009) 5550–5555. https://doi.org/10.1021/ef900597q.

[2] T.M. Lovestead, B.C. Windom, J.R. Riggs, C. Nickell, T.J. Bruno, Assessment of the compositional variability of RP-1 and RP-2 with the advanced distillation curve approach, Energy Fuels. 24 (2010) 5611–5623. https://doi.org/10.1021/ef100994w.

[3] T.J. Bruno, B.C. Windom, Method and apparatus for the thermal stress of complex fluids: Application to fuels, Energy Fuels. 25 (2011) 2625–2632. https://doi.org/10.1021/ef2004738.

[4] T.M. Lovestead, J.L. Burger, N. Schneider, T.J. Bruno, Comprehensive Assessment of Composition and Thermochemical Variability by High Resolution GC/QToF-MS and the Advanced Distillation-Curve Method as a Basis of Comparison for Reference Fuel Development, Energy Fuels. 30 (2016) 10029–10044. https://doi.org/10.1021/acs.energyfuels.6b01837.

[5] K.L. Berrier, C.E. Freye, M.C. Billingsley, R.E. Synovec, Predictive Modeling of Aerospace Fuel Properties Using Comprehensive Two-Dimensional Gas Chromatography with Time-Of-Flight Mass Spectrometry and Partial Least Squares Analysis, Energy Fuels. 34 (2020) 4084–4094. https://doi.org/10.1021/acs.energyfuels.9b04108.

[6] D.K. Huzel, D.H. Huang, Modern Engineering for Design of Liquid-Propellant Rocket Engines, American Institute of Aeronautics and Astronautics, Inc., Washington, DC, 1992.

[7] O. Altin, S. Eser, Analysis of Carboneceous Deposits from Thermal Stressing of a JP-8 Fuel on Superalloy Foils in a Flow Reactor, Ind. Eng. Chem. Res. 40 (2001) 589–595. https://doi.org/10.1021/ie0004489.

[8] S. Eser, R. Venkataraman, O. Altin, Deposition of carbonaceous solids on different substrates from thermal stressing of JP-8 and jet A fuels, Ind. Eng. Chem. Res. 45 (2006) 8946–8955. https://doi.org/10.1021/ie060968p.

[9] L. Guozhu, H. Yongjin, W. Li, Z. Xiangwen, M. Zhentao, Solid deposits from thermal stressing of n - Dodecane and chinese RP-3 jet fuel in the presence of several initiators, Energy Fuels. 23 (2009) 356–365. https://doi.org/10.1021/ef800657z.

[10] P.C. Andersen, T.J. Bruno, Thermal decomposition kinetics of RP-1 rocket propellant, Ind. Eng. Chem. Res. 44 (2005) 1670–1676. https://doi.org/10.1021/ie048958g.

[11] T.J. Bruno, J.A. Widegren, Thermal decomposition kinetics of kerosene-based rocket propellants. 1. comparison of RP-1 and RP-2, Energy Fuels. 23 (2009) 5517–5522. https://doi.org/10.1021/ef900576g.

[12] T.J. Bruno, J.A. Widegren, Thermal decomposition kinetics of kerosene-based rocket propellants. 2. RP-2 with three additives, Energy Fuels. 23 (2009) 5523–5528. https://doi.org/10.1021/ef900577k.

[13] T.J. Fortin, T.J. Bruno, Assessment of the thermophysical properties of thermally stressed RP-1 and RP-2, Energy Fuels. 27 (2013) 2506–2514. https://doi.org/10.1021/ef400193d.

[14] N.J. Kuprowicz, S. Zabarnick, Z.J. West, J.S. Ervin, Use of measured species class concentrations with chemical kinetic modeling for the prediction of autoxidation and deposition of jet fuels, Energy Fuels. 21 (2007) 530–544. https://doi.org/10.1021/ef060391o.

[15] M. Sobkowiak, J.M. Griffith, B. Wang, B. Beaver, Insight into the mechanisms of middle distillate fuel oxidative degradation. Part 1: On the role of phenol, indole, and carbazole derivatives in the thermal oxidative stability of fischer-tropsch/petroleum jet fuel blends, Energy Fuels. 23 (2009) 2041–2046. https://doi.org/10.1021/ef8006992.

[16] E. Alborzi, P. Gadsby, M.S. Ismail, A. Sheikhansari, M.R. Dwyer, A.J.H.M. Meijer, S.G. Blakey, M. Pourkashanian, Comparative Study of the Effect of Fuel Deoxygenation and Polar Species Removal on Jet Fuel Surface Deposition, Energy Fuels. 33 (2019) 1825–1836. https://doi.org/10.1021/acs.energyfuels.8b03468.

[17] S. Zabarnick, Z.J. West, L.M. Shafer, S.S. Mueller, R.C. Striebich, P.J. Wrzesinski, Studies of the Role of Heteroatomic Species in Jet Fuel Thermal Stability: Model Fuel Mixtures and Real Fuels, Energy Fuels. 33 (2019) 8557–8565. https://doi.org/10.1021/acs.energyfuels.9b02345.

[18] N.J. Begue, J.A. Cramer, C. Von Bargen, K.M. Myers, K.J. Johnson, R.E. Morris, Automated method for determining hydrocarbon distributions in mobility fuels, Energy Fuels. 25 (2011) 1617–1623. https://doi.org/10.1021/ef101635a.

[19] R. V. Gough, T.J. Bruno, Composition-explicit distillation curves of alternative turbine fuels, Energy Fuels. 27 (2013) 294–302. https://doi.org/10.1021/ef3016848.

[20] P.Y. Hsieh, K.R. Abel, T.J. Bruno, Analysis of marine diesel fuel with the advanced distillation curve method, Energy Fuels. 27 (2013) 804–810. https://doi.org/10.1021/ef3020525.

[21] P.E. Sudol, K.M. Pierce, S.E. Prebihalo, K.J. Skogerboe, B.W. Wright, R.E. Synovec, Development of gas chromatographic pattern recognition and classification tools for compliance and forensic analyses of fuels: A review, Anal. Chim. Acta. 1132 (2020) 157–186. https://doi.org/10.1016/j.aca.2020.07.027.

[22] Z. Liu, J.B. Phillips, Comprehensive two-dimensional gas chromatography using an on-column thermal modulator interface, J. Chromatogr. Sci. 29 (1991) 227–231. https://doi.org/10.1093/chromsci/29.6.227.

[23] M.S. Klee, J. Cochran, M. Merrick, L.M. Blumberg, Evaluation of conditions of comprehensive two-dimensional gas chromatography that yield a near-theoretical maximum in peak capacity gain, J. Chromatogr. A. 1383 (2015) 151–159. https://doi.org/10.1016/j.chroma.2015.01.031.

[24] Z. Liu, J.B. Phillips, Sensitivity and detection limit enhancement of gas chromatographic detection by thermal modulation, J. Microcolumn Sep. 6 (1994) 229–235. https://doi.org/10.1002/mcs.1220060306.

[25] A.L. Lee, K.D. Bartle, A.C. Lewis, A model of peak amplitude enhancement in orthogonal two-dimensional gas chromatography, Anal. Chem. 73 (2001) 1330–1335. https://doi.org/10.1021/ac001120s.

[26] M.K. Jennerwein, M. Eschner, T. Gröger, T. Wilharm, R. Zimmermann, Complete group-type quantification of petroleum middle distillates based on comprehensive two-dimensional gas chromatography time-of-flight mass spectrometry (GCxGC-TOFMS) and visual basic scripting, Energy Fuels. 28 (2014) 5670–5681. https://doi.org/10.1021/ef501247h.

[27] B. Kehimkar, J.C. Hoggard, L.C. Marney, M.C. Billingsley, C.G. Fraga, T.J. Bruno, R.E. Synovec, Correlation of rocket propulsion fuel properties with chemical composition using comprehensive two-dimensional gas chromatography with time-of-flight mass spectrometry followed by partial least squares regression analysis, J. Chromatogr. A. 1327 (2014) 132–140. https://doi.org/10.1016/j.chroma.2013.12.060.

[28] B. Kehimkar, B.A. Parsons, J.C. Hoggard, M.C. Billingsley, T.J. Bruno, R.E. Synovec, Modeling RP-1 fuel advanced distillation data using comprehensive two-dimensional gas chromatography coupled with time-of-flight mass spectrometry and partial least squares analysis, Anal. Bioanal. Chem. 407 (2015) 321–330. https://doi.org/10.1007/s00216-014-8233-6.

[29] C.E. Freye, B.D. Fitz, M.C. Billingsley, R.E. Synovec, Partial least squares analysis of rocket propulsion fuel data using diaphragm valve-based comprehensive two-dimensional gas chromatography coupled with flame ionization detection, Talanta. 153 (2016) 203–210. https://doi.org/10.1016/j.talanta.2016.03.016.

[30] V. Abrahamsson, N. Ristic, K. Franz, K. Van Geem, Comprehensive two-dimensional gas chromatography in combination with pixel-based analysis for fouling tendency prediction, J. Chromatogr. A. 1501 (2017) 89–98. https://doi.org/10.1016/j.chroma.2017.04.021.

[31] M.K. Jennerwein, A.C. Sutherland, M. Eschner, T. Gröger, T. Wilharm, R. Zimmermann, Quantitative analysis of modern fuels derived from middle distillates – The impact of diverse compositions on standard methods evaluated by an offline hyphenation of HPLC-refractive index detection with GC×GC-TOFMS, Fuel. 187 (2017) 16–25. https://doi.org/10.1016/j.fuel.2016.09.033.

[32] X. Shi, H. Li, Z. Song, X. Zhang, G. Liu, Quantitative composition-property relationship of aviation hydrocarbon fuel based on comprehensive two-dimensional gas chromatography with mass spectrometry and flame ionization detector, Fuel. 200 (2017) 395–406. https://doi.org/10.1016/j.fuel.2017.03.073.

[33] R.L. Webster, P.M. Rawson, C. Kulsing, D.J. Evans, P.J. Marriott, Investigation of the Thermal Oxidation of Conventional and Alternate Aviation Fuels with Comprehensive Two-Dimensional Gas Chromatography Accurate Mass Quadrupole Time-of-Flight Mass Spectrometry, Energy Fuels. 31 (2017) 4886–4894. https://doi.org/10.1021/acs.energyfuels.7b00178.

[34] R. Chakravarthy, C. Acharya, A. Savalia, G.N. Naik, A.K. Das, C. Saravanan, A. Verma, K.B. Gudasi, Property Prediction of Diesel Fuel Based on the Composition Analysis Data

by two-Dimensional Gas Chromatography, Energy Fuels. 32 (2018) 3760–3774. https://doi.org/10.1021/acs.energyfuels.7b03822.

[35] P. Vozka, H. Mo, P. Šimáček, G. Kilaz, Middle distillates hydrogen content via GC×GC-FID, Talanta. 186 (2018) 140–146. https://doi.org/10.1016/j.talanta.2018.04.059.

[36] P. Vozka, B.A. Modereger, A.C. Park, W.T.J. Zhang, R.W. Trice, H.I. Kenttämaa, G. Kilaz, Jet fuel density via GC × GC-FID, Fuel. 235 (2019) 1052–1060. https://doi.org/10.1016/j.fuel.2018.08.110.

[37] J. Heyne, D. Bell, J. Feldhausen, Z. Yang, R. Boehm, Towards fuel composition and properties from Two-dimensional gas chromatography with flame ionization and vacuum ultraviolet spectroscopy, Fuel. 312 (2022) 122709. https://doi.org/10.1016/j.fuel.2021.122709.

[38] G.S. Ochoa, M.C. Billingsley, R.E. Synovec, Using solid-phase extraction to facilitate a focused tile-based Fisher ratio analysis of comprehensive two-dimensional gas chromatography time-of-flight mass spectrometry data: comparative analysis of aerospace fuel composition, Anal. Bioanal. Chem. (2022). https://doi.org/10.1007/s00216-022-04348-1.

[39] M. Jennerwein, M. Eschner, T. Wilharm, T. Gröger, R. Zimmermann, Evaluation of reversed phase versus normal phase column combination for the quantitative analysis of common commercial available middle distillates using GC × GC-TOFMS and Visual Basic Script, Fuel. 235 (2019) 336–338. https://doi.org/10.1016/j.fuel.2018.07.081.

[40] T.J. Trinklein, C.N. Cain, G.S. Ochoa, S. Schöneich, L. Mikaliunaite, R.E. Synovec, Recent Advances in GC×GC and Chemometrics to Address Emerging Challenges in Nontargeted Analysis, Anal. Chem. 95 (2023) 264–286. https://doi.org/10.1021/acs.analchem.2c04235.

[41] C.N. Cain, G.S. Ochoa, R.E. Synovec, Enhancing partial least squares modeling of comprehensive two-dimensional gas chromatography time-of-flight mass spectrometry data by tile-based variance ranking, J. Chromatogr. A. 1694 (2023) 463920. https://doi.org/10.1016/j.chroma.2023.463920.

[42] T.J. Trinklein, J. Jiang, R.E. Synovec, Profiling Olefins in Gasoline by Bromination Using GC×GC-TOFMS Followed by Discovery-Based Comparative Analysis, Anal. Chem. 94 (2022) 9407–9414. https://doi.org/10.1021/acs.analchem.2c01549.

[43] L.C. Marney, W.C. Siegler, B.A. Parsons, J.C. Hoggard, B.W. Wright, R.E. Synovec, Tile-based Fisher-ratio software for improved feature selection analysis of comprehensive two-dimensional gas chromatography-time-of-flight mass spectrometry data, Talanta. 115 (2013) 887–895. https://doi.org/10.1016/j.talanta.2013.06.038.

[44] B.A. Parsons, L.C. Marney, W.C. Siegler, J.C. Hoggard, B.W. Wright, R.E. Synovec, Tile-Based Fisher Ratio Analysis of Comprehensive Two-Dimensional Gas Chromatography Time-of-Flight Mass Spectrometry (GC × GC–TOFMS) Data Using a Null Distribution Approach, Anal. Chem. 87 (2015) 3812–3819. https://doi.org/10.1021/ac504472s.

[45] S. Wold, K. Esbensen, P. Geladi, Principal component analysis, Chemom. Intell. Lab. Syst. 2 (1987) 37–52. https://doi.org/10.1016/0169-7439(87)80084-9.

[46] R.E. Synovec, C.E. Freye, B.A. Parsons, M.C. Billingsley, N. Keim, B. Hill-Lam, J.C. Wilhelm, Recent Advances in the Development of Chemical Analysis Tools to Relate Compositional Variation to Thermal Integrity Data for RP-1, RP-2, and Related Fuels, in: JANNAF 8th Liq. Propuls. Meet., Nashville, TN, 2015.

[47] R.E. Synovec, C.E. Freye, M.C. Billingsley, N. Keim, B. Hill-Lam, Recent Advances in the Development of Chemical Analysis Tools to Relate Compositional Variation to Thermal Integrity Data for RP-1, RP-2, and Related Fuels, in: JANNAF 9th Liq. Propuls. Meet., Phoenix, AZ, 2016.

[48] R.E. Synovec, C.E. Freye, M.C. Billingsley, N. Keim, B. Hill-Lam, Recent Advances in Relating Chemical Compositional Variation in RP-1, RP-2, and Similar Fuels to Thermal Integrity Data, in: JANNAF 10th Liq. Propuls. Meet. Long Beach, CA, 2018.

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

[50] G.S. Ochoa, S.E. Prebihalo, B.C. Reaser, L.C. Marney, R.E. Synovec, Statistical inference of mass channel purity from Fisher ratio analysis using comprehensive two-dimensional gas chromatography with time of flight mass spectrometry data, J. Chromatogr. A. 1627 (2020) 461401. https://doi.org/10.1016/j.chroma.2020.461401.

[51] P.E. Sudol, D. V. Gough, S.E. Prebihalo, R.E. Synovec, Impact of data bin size on the classification of diesel fuels using comprehensive two-dimensional gas chromatography with principal component analysis, Talanta. 206 (2020) 120239. https://doi.org/10.1016/j.talanta.2019.120239.

[52] J.W. Frankenfeld, W.F. Taylor, Deposit Formation from Deoxygenated Hydrocarbons. 4. Studies in Pure Compound Systems, Ind. Eng. Chem. Prod. Res. Dev. 19 (1980) 65–70. https://doi.org/10.1021/i360073a016.

[53] M.J. Dewitt, T. Edwards, L. Shafer, D. Brooks, R. Striebich, S.P. Bagley, M.J. Wornat, Effect of aviation fuel type on pyrolytic reactivity and deposition propensity under supercritical conditions, Ind. Eng. Chem. Res. 50 (2011) 10434–10451. https://doi.org/10.1021/ie200257b.

[54] Z. Fan, P. Rahimi, T. Alem, A. Eisenhawer, P. Arboleda, Fouling characteristics of hydrocarbon streams containing olefins and conjugated olefins, Energy Fuels. 25 (2011) 1182–1190. https://doi.org/10.1021/ef101661n.

[55] M. Gao, L. Hou, X. Zhang, D. Zhang, Coke deposition inhibition for endothermic hydrocarbon fuels in a reforming catalyst-coated reactor, Energy Fuels. 33 (2019) 6126–6133. https://doi.org/10.1021/acs.energyfuels.9b00878.

[56] M. Commodo, O. Wong, I. Fabris, C.P.T. Groth, Ö.L. Gülder, Spectroscopic study of aviation jet fuel thermal oxidative stability, Energy Fuels. 24 (2010) 6437–6441. https://doi.org/10.1021/ef1012837.

[57] A.R. Mohan, S. Eser, Analysis of carbonaceous solid deposits from thermal oxidative stressing of Jet-A fuel on iron- and nickel-based alloy surfaces, Ind. Eng. Chem. Res. 49 (2010) 2722–2730. https://doi.org/10.1021/ie901283r.

[58] J. Yu, S. Eser, Thermal Decomposition of C10-C14 Normal Alkanes in Near-Critical and Supercritical Regions: Product Distributions and Reaction Mechanisms, Ind. Eng. Chem. Res. 36 (1997) 574–584. https://doi.org/10.1021/ie960392b.

[59] J.A. Widegren, T.J. Bruno, Thermal stability of RP-2 as a function of composition: The effect of linear, branched, and cyclic alkanes, Energy Fuels. 27 (2013) 5138–5143. https://doi.org/10.1021/ef401677g.

Chapter 5: Development of Variance Rank Initiated-Unsupervised Sample Indexing for Gas Chromatography-Mass Spectrometry Analysis

5.1. Introduction

Gas chromatography-mass spectrometry (GC-MS) is a popular analytical tool for the analysis of volatile and semi-volatile mixtures throughout many fields, including metabolomics, environmental, petrochemical, and food samples [1–4]. The temporal and spectral resolution obtained by GC-MS provides an ideal data set for chemometric methods, which seek to discover meaningful chemical information using mathematical means. Chemometric techniques are commonly utilized in GC-MS analyses for data reduction through feature selection [5–9], categorization of samples based on similarities/differences [10–13], and property prediction [14,15]. Depending on the analysis goals, these algorithms can be applied in either a targeted or non-targeted approach. To define, targeted methods aim to evaluate previous classification or prediction models using specific marker analytes, while non-targeted methods focus on the discovery of novel compounds that define the data set [8]. The latter approach is the most promising for exploratory investigations into a data set [16].

Non-targeted chemometric techniques can be categorized as either supervised or unsupervised, depending upon whether class membership is known a priori. Supervised methods use target variable(s), such as class membership or independently measured properties, to determine their relationship to the data set [17]. Commonly, supervised data analysis workflows will implement a feature selection method, which select a subset of the original variables that are

---

This chapter is reproduced from C. N. Cain*, P.E. Sudol*, K. L. Berrier*, R. E. Synovec, Development of Variance Rank Initiated-Unsupervised Sample Indexing for Gas Chromatography-Mass Spectrometry Analysis, Talanta 233 (2021), 122495. * These authors contributed equally.

highly correlated to a target variable, prior to developing classification or regression models [5,6,10,12–14,17,18]. In contrast, unsupervised chemometric methods discover patterns or outliers in a data set without prior knowledge of target variable(s) [17,19]. The increased use of unsupervised methods is due to a recent push to further understand the underlying nature of the data, detect sample heterogeneities, and eliminate human biases in data analysis [20].

Unsupervised chemometric techniques can be subdivided into dimensionality reduction or clustering methods. Given the high dimensionality of a GC-MS data set, it could be beneficial to reduce the data set down to a few chromatographic variables that best describe the overall variation in the data set [17]. Principal component analysis (PCA) is a powerful dimensionality reduction method, finding the linear combination of variables that maximizes the overall variance [21]. PCA is commonly implemented in the exploratory analysis of GC-MS data since the similarities/differences between samples can be easily visualized on scores plots [5,6,11,13,17,22]. While PCA can visualize the differences between samples, clustering algorithms explicitly identify the natural groupings of samples in a data set [17,20]. Although numerous clustering methods have been reported, k-means clustering is the most popular due to its ease of use, simplicity, and efficiency [17,20]. This clustering algorithm will assign samples into a specified number of clusters, *k*, in order to reduce the sum-of-square error between cluster centroids [17,20]. *K*-means clustering has been used in GC-MS studies to characterize metabolomics, environmental, and food samples [23–26].

Even though unsupervised methods are beneficial in exploratory data analysis, interpretation and validation of the quality of results ultimately relies upon a priori knowledge [27,28]. Herein, we present a chemometric workflow for GC-MS data, termed variance rank initiated-unsupervised sample indexing (VRI-USI), to address this issue for unsupervised

methods. VRI-USI first calculates the relative variance of each peak feature across all samples and ranks the peaks accordingly to identify variables with high variance and thus potentially high predictive power [18,29–32]. While *k*-means clustering can be directly applied to chromatographic data, data reduction with PCA has been shown to improve the effectiveness of clustering [33]. Therefore, VRI-USI performs *k*-means clustering to the resulting PCA scores plot developed for each discovered feature [33]. Based upon how the samples cluster for a given discovered peak feature, a sample index assignment is provided, which is essentially the groupings of the sample numbers in each cluster. To determine the correct "underlying" sample clustering for the entire data set and ease interpretation of the results, the sample index assignments for each discovered peak are compared to each another. If the same sample index assignment appears for several discovered peaks, then there is a high probability of samples being properly classified by that specific sample index assignment, especially if these indices are observed for analytes with a large relative variance. Application of VRI-USI is demonstrated on simulated GC-MS data in a pixel-based approach and implemented on peak tables from yeast metabolome [34–36] and human cancer data sets [37].

5.2. Theory

For an analytical data set, the total relative signal variance, $V_{\mathrm{R,T}}$, can be expressed as

$$V_{\mathrm{R,T}} = V_{\mathrm{R,Chem}} + V_{\mathrm{R,B}} \tag{5.1}$$

where $V_{\mathrm{R,Chem}}$ is the chemically significant relative signal variance and $V_{\mathrm{R,B}}$ is the "background" relative signal variance. To clarify, relative variance is defined here as the square of the standard deviation normalized by the square of the mean to obtain a dimensionless quantity equivalent to the square of the relative standard deviation (RSD^2). The background variance (RSD^2_{B}) can be expressed as follows,

$$RSD^2{}_{\mathrm{B}} = RSD^2{}_{\mathrm{NatVar}} + RSD^2{}_{\mathrm{SP}} + RSD^2{}_{\mathrm{Inj}} + RSD^2{}_{\mathrm{Det}} \quad (5.2)$$

with subscripts corresponding to the four major sources of uncertainty in experiments: natural variation (NatVar), such as biological variation, sample preparation (SP), injection (Inj), and detection (Det). Here, all sources of background variation are the common sources of variation that are present in GC-MS measurements. The first variance contribution, RSD^2_{NatVar}, is often the dominating term for bioanalytical studies such as metabolomics. For the yeast metabolome data set used herein, the RSD^2_{NatVar} ranged from 0.09 – 0.25 (30 – 50 % *RSD*) [34]. The second contribution, RSD^2_{SP}, typically ranges of 0.01 – 0.04 (10 – 20 % *RSD*) for solvent extraction-based sample preparation but can be larger for solid-phase microextraction (SPME) [38,39]. The third contribution, RSD^2_{Inj}, is due to sample volume injection variation with GC-based instrumentation, which can range from 0.0025 – 0.01 (5 – 10 % *RSD*) [40]. The use of data normalization or internal standards can minimize the contributions from both RSD^2_{SP} and RSD^2_{Inj}, ensuring these contributions do not inflate the $V_{\mathrm{R,T.}}$ The final contribution, RSD^2_{Det}, is defined in terms of the limit-of-detection (*LOD*), with the signal at the *LOD* equal to $3\times\sigma_{\mathrm{N}}$, whereby σ_{N} is the standard deviation of the baseline noise. The magnitude of RSD^2_{Det} will be small for analyte signals with a sufficiently high signal-to-noise ratio (*S/N*). Since analyte peaks with low *S/N* could inadvertently have inflated $V_{\mathrm{R,T}}$ from the RSD^2_{Det} contribution, applying a suitable signal threshold will limit VRI-USI to only discover peaks with a sufficient *S/N*. Therefore, peaks with a larger total RSD^2 are more likely to have additional source(s) of signal variance ($V_{\mathrm{R,Chem}}$), thus more potential to discover useful chemical information in the data set.

After the peaks are discovered and their relative variances are ranked, the *k*-means clustering output provides sample clustering information, i.e., the sample index assignments. For a given number of clusters (*k*), the number of unique combinations (*N*) of samples in each cluster is

$$N = \binom{s}{r_1, r_2, \ldots, r_k} = \frac{s!}{r_1! r_2! \ldots r_k!} \tag{5.3}$$

where s is the number of samples in the data set and $\{r_1, r_2, \ldots, r_k\}$ is the number of samples assigned into each cluster [41]. The probability of one specific sample index assignment (p_k) occurring is calculated as the inverse of N because only one of the unique combinations will be the given sample index assignment. If x number of analyte peaks have the same sample index assignment from k-means clustering, then the probability can be calculated as a binomial probability (P_x) [41]:

$$P_x = \binom{n}{x} {p_k}^x (1 - p_k)^{n-x} \tag{5.4}$$

where n is the number of peaks discovered in the initial relative variance ranking step. The more peaks discovered to have the same sample index assignments will lead to a smaller probability of those sample cluster indices occurring by chance, and conversely, a higher probability that the sample index assignments are indicating the samples are properly classified by that particular set of sample index assignments. Thus, an analyst can gain confidence that meaningful chemical differences among the samples are causing those sample index assignments.

5.3. Methods and Materials

5.3.1. Chromatographic simulations

All simulations were performed in Matlab R2020a (Mathworks, Inc., Natick, MA, U.S.A.), where each simulation consisted of 10 replicates of two chemically distinct classes, Class A and Class B, to give 20 sample replicates total. Two data sets, each containing 50 chromatographic simulations, were simulated using the same retention times, peak heights, mass spectrum library, and class-indicating analytes. The only difference between these two chromatographic data sets is the $V_{R,B}$ conditions simulated. For the first 50 chromatographic data sets, a $V_{R,B}$ of 0.09 (30 % *RSD*) was randomly generated for each analyte. Another 50

chromatographic data sets were simulated to have the $V_{R,B}$ of each analyte vary between 0.09 and 0.25 (30 – 50 % *RSD*). The additional simulation parameters are summarized below and in Table C.1.

Chromatograms were simulated to contain 50 analytes randomly and independently distributed throughout a separation window of 50 s. Each analyte was simulated as a Gaussian peak with a constant width-at-base (W_b, $\pm 2\sigma$) of 1 s at a mass spectral scan rate of 10 Hz (1 data point = 100 ms). Analogous to the peak height distributions observed experimentally [42,43], an exponential distribution in peak heights with a mean of 100 was randomly and independently generated for each simulation. Analyte concentrations in Class B remained nominally the same while concentrations (i.e., peak heights) of the six randomly selected analytes in Class A were changed by factors of 0.33, 0.5, 0.67, 1.5, 2, and 3 with respect to the concentrations in Class B. A randomly selected mass spectrum obtained from the NIST MS Search 2.0 database (NJ, U.S.A.) was then multiplied elementwise across the peak. The mass spectrum of each analyte was normalized such that the sum of the intensities of all mass channels, *m/z*, would be equal to 1000. Random Gaussian-distributed noise was generated independently for each *m/z* with a standard deviation that would provide the desired *S/N* of 20 for the mean peak height in the TIC chromatogram (~100,000). The actual *S/N* for each analyte depended on the exponentially distributed peak height, and the *S/N* for each *m/z* depended on the intensity of the *m/z*.

5.3.2. Yeast metabolome data set

Separations of two classes of yeast, repressed (R) and derepressed (DR) were previously collected using GC×GC-TOFMS [34,35]. The repressed class metabolized glucose to enact fermentation and the derepressed class was provided with ethanol to cause respiration. Three yeast cultures for each class were maintained (A, B, C), followed by three extractions of each

culture, and four injection replicates. Experimental details on the sample preparation and chromatographic separation of the yeast cell metabolite extracts can be found in the original articles [34,35] and in Appendix C. For the present study, six samples for each class were selected so that two samples from each culture were present for each class (Table C.2). Only one extraction replicate for each culture is represented in each class to minimize the retention time shifting observed in the original data set [35,36]. Also, previous work indicated that any combination of extraction replicates yielded similar results in the discovery of class-distinguishing and false positive metabolites [36].

The sample data files were imported into Matlab using an in-house algorithm, where the data were baseline corrected, centered with the mean of the baseline around 0, and normalized to the average TIC signal. Data collected prior to 10 min was removed due to a lack of peaks. Mass channel 44 was zero-filled due to a large background noise stemming from the presence of CO_2. For this proof-of-principle study, the GC×GC-TOFMS data set was reduced to a one-dimensional GC-MS format by summing the second dimension, ^{2}D, onto the first dimension, ^{1}D. This resulted in data in a GC-MS format whereby the data collection rate was reduced to 0.67 Hz (1 data point = 1 modulation = 1.5 s). While this is a low effective sampling rate, the data provided were suitable for the purposes of this study. Due to minor retention time shifting on the ^{1}D dimension (± 1 modulation), a peak table was assembled. First, a peak finder was employed on the data in the GC-MS format for each *m/z* and sample to locate the time associated with peak maxima. Using an in-house peak table alignment algorithm, the detected peaks were aligned across all samples and *m/z*. A deviation of ± 1 modulation with a maximum range in retention time of 2 modulations was allowed. Peaks included in the finalized table were required to have at

least five samples (i.e., one less than half of the samples) pass a signal threshold of 10-fold the *LOD* and be defined by at least three *m/z*. A total of 53 peaks met these criteria.

5.3.3. Head and neck cancer metabolomic data set

A previously published data set of volatile metabolites in human saliva [37] was downloaded from MetaboLights (Study Identifier MTBLS760) [44]. Headspace-SPME-GC-MS separations were collected for 32 head and neck cancer and 27 control patients [37]. Experimental details regarding sample collection, preparation, and analysis can be found in the original article [37] and in Appendix C. A table of peak areas for 48 volatile analytes was assembled for analysis since they were confidently identified (greater than 75 % match with a NIST library) and did not contain more than 50 % missing values [37]. The peak table was imported into Matlab, where missing values were replaced by half of the minimum positive value [37] and the data were median normalized.

5.3.4. Variance rank initiated-unsupervised sample indexing (VRI-USI)

For the two simulated data sets, the RSD^2 was calculated for each data point that had at least one sample above the signal threshold of 10-fold the *LOD* across all *m/z*, which was determined by finding the greatest σ_N across all *m/z* in a user-defined baseline region. The average RSD^2 for each time point was calculated using the top 10 *m/z* with the largest RSD^2 (a minimum of 3 *m/z* were required to be above the signal threshold). The local maxima in RSD^2 were discovered, ranked in descending order, and arranged into "hit lists". Since the yeast and human metabolomics data sets were already filtered by the signal threshold, the RSD^2 was directly calculated for the signals in the peak tables and those tables were arranged to rank RSD^2 in descending order. Next, peak profiles for the simulated and yeast metabolome data sets, defined by the W_b for each peak, were created by summing together all the *m/z* that defined that

peak. PCA was performed on the resulting mean-centered peak signal vectors (simulated and yeast data sets) and scalar peak areas (human metabolome data set) using the PLS Toolbox version 8.8.1 (Eigenvector Research Inc., Wenatchee, WA, USA). Using the scores on the first two PCs, *k*-means was performed to cluster the samples into two groups ($k = 2$) and three groups ($k = 3$) [20,33]. The Manhattan distance, which calculates the sum of absolute differences, was utilized as the distance metric for *k*-means clustering [45]. Ten iterations of *k*-means clustering for each hit at $k = 2$ and $k = 3$ were performed to find the cluster centroid positions that minimizes the sum of distances. The sample index assignments designated by *k*-means were then added to the hit lists and the hit lists were examined to discover peaks with matching sample index assignments. Using the sample index assignments at $k = 2$ with the lowest probability of occurring by chance, a concentration ratio and results from a Welch's *t*-test (i.e., unequal variances *t*-test) at the 95 % confidence interval was reported for each peak in the hit lists.

5.4. Results and Discussion

5.4.1. Evaluation of VRI-USI with chromatographic simulations

Chromatographic replicates of two sample classes were simulated to demonstrate VRI-USI in a controlled setting where the background variation and locations of class-indicating analytes are known. Figure 5.1A-B shows the TIC chromatograms of Class A (red) and Class B (blue) for simulation #13, representing the two types of chromatographic data sets. Figure 5.1A illustrates a simulation where the $V_{R,B}$ was constant at 0.09 (30 % RSD_B), while Figure 5.1B shows that the $V_{R,B}$ for each analyte could vary between 0.09 and 0.25 (30 – 50 % RSD_B), which is akin to natural variation for biological samples [34]. This simulation was arbitrarily chosen for illustrative purposes. The peak heights of fifty analytes are simulated using an exponential distribution to also mimic natural samples [42,43]. Due to the exponential distribution of peak

heights and peak crowdedness ($\alpha = 1$) modeled, all fifty analytes may not be readily observed in the TIC chromatograms. The RSD^2 was calculated at each time point per-*m/z* in the simulated data matrix and the top 10 *m/z* with the highest RSD^2 were averaged together. Figure 5.1C-D shows the resulting RSD^2 metric as a function of retention time for chromatographic simulations shown in Figure 5.1A-B. Since Figure 5.1A was simulated to have a constant $V_{R,B}$ of 0.09, Figure 5.1C shows that a majority of these signals have an RSD^2 equal to 0.09 except for six regions of signal with a RSD^2 greater than 0.09. These six regions correspond to the six chromatographic peaks that contain additional variance from $V_{R,Chem}$, due to the simulated concentration changes. Likewise, since the $V_{R,B}$ of each analyte in Figure 5.1B could vary between 0.09 and 0.25, Figure 5.1D shows the difficulty in identifying chromatographic signals that have the additional contribution of $V_{R,Chem}$. Regardless of if the peaks with the additional $V_{R,Chem}$ contribution can be distinguished in Figure 5.1C-D, application of VRI-USI to these simulations will allow for identification of both the class-indicating peaks with the concentration changes, and thus, the sample groupings.

Using the calculated RSD^2 metric, hit lists were obtained for the simulations shown in Figure 5.1. Starting with the constant $V_{R,B}$ simulation (Figure 5.1A), Table 5.1 shows the first 12 peaks in the hit list, ranked by RSD^2, while the entire hit list is shown in Table C.3. From the hit list, the first six hits all have the highest RSD^2 values while the remaining 20 hits all have RSD^2 approximately equal to the $V_{R,B}$ simulated. For each peak in the hit list, the peak profile was imported into PCA and *k*-means clustering at $k = 2$ was applied to the scores plot to determine the sample index assignments. Table 5.1 shows that the first six hits all have matching index assignments (shaded in green) while the remaining hits had inconsistent index assignments when compared (Table C.3). Applying Eqs. 5.3 and 5.4, the probability of having six matching index

assignments out of 26 total hits occurring by chance is 5.89×10^{-27}, indicating that it is very unlikely that these matches occurred by chance. Additionally, Table C.4 illustrates that none of the hits have matching sample index assignments after performing *k*-means clustering at $k = 3$. Therefore, these results indicate that there is an underlying sample-to-sample relationship that is causing the 20 samples to cluster into these two specific groups for the first six hits. Assuming these sample index assignments define two sample classes for all peaks in the hit list (Table 5.1), a concentration ratio and *p*-value was calculated. Examination of the hit list reveals that the first six peaks have a statistically significant ($p < 0.05$) concentration ratio matching the changes simulated in the peak heights at those retention times (Figure 5.1A). Peaks that had an RSD^2 equal to 0.09 had concentration ratios approximately equal to 1 and were not statistically significant ($p > 0.05$) because these peaks were simulated to only have a $V_{R,B}$ contribution.

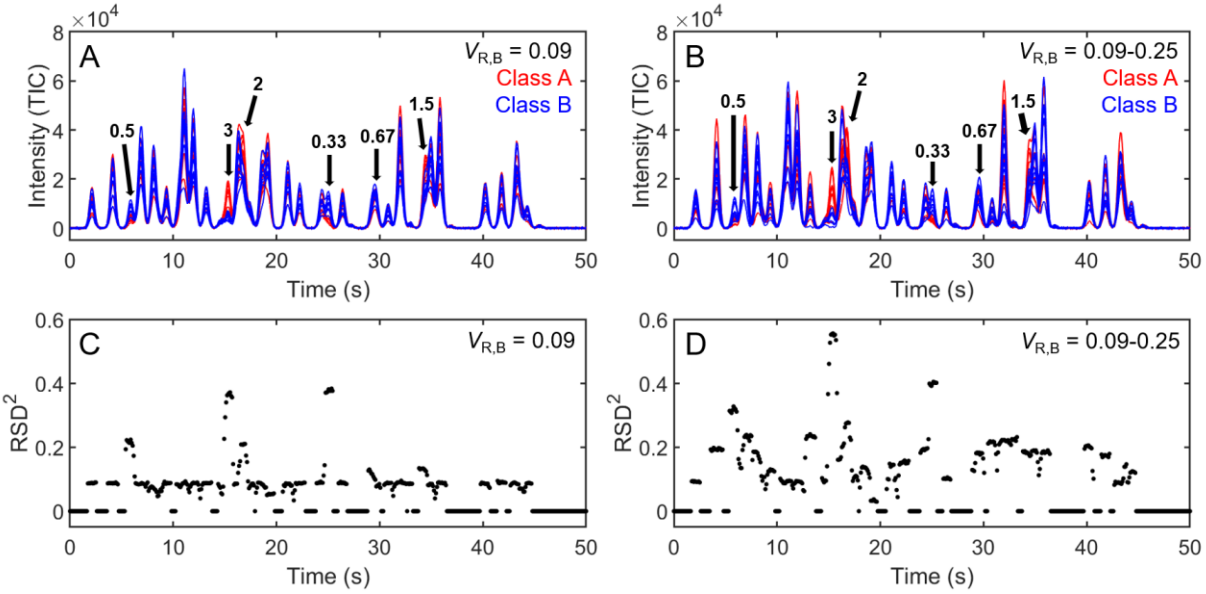


Figure 5.1. Calculation of relative signal variance for chromatographic simulations containing a background variance ($V_{R,B}$) of 0.09 (30 % *RSD*) (A,C), and a background variance ($V_{R,B}$) of 0.09 – 0.25 (30 – 50 % *RSD*) (B,D). The data set shown is from simulation #13 from each set of 50 simulations. (A,B) Overlaid TIC chromatograms of Class A (red) and Class B (blue) replicates. Black arrows indicate the six analytes selected to be upregulated or downregulated in Class A. (C,D) RSD^2 calculated for the respective simulated data sets as a function of retention time.

Each black dot in the RSD^2 plot represents the average of the RSD^2 for top 10 *m/z* at that time point.

Table 5.1. Resulting hit list after application of VRI-USI to the simulated data set containing a background variance of 0.09 (30 % *RSD*), which is shown in Figure 5.1A. The top 12 hits, ranked by RSD^2, are shown for brevity.[a]

Hit Number	t_R (s)	RSD^2	Sample Index Assignments	[Class A]/[Class B]	*p*-value
1	25.34	0.38	Samples 1-10; Samples 11-20	0.33	< 0.001
2	15.51	0.37	Samples 1-10; Samples 11-20	2.99	< 0.001
3	5.88	0.22	Samples 1-10; Samples 11-20	0.50	< 0.001
4	16.96	0.21	Samples 1-10; Samples 11-20	2.00	0.001
5	33.82	0.13	Samples 1-10; Samples 11-20	1.51	0.009
6	29.00	0.13	Samples 1-10; Samples 11-20	0.67	0.009
7	32.08	0.10	Samples 1,4,6-10; Samples 2,3,4,11-20	1.00	0.993
8	41.43	0.09	Samples 1-7; Samples 8-20	1.00	0.985
9	11.18	0.09	Samples 1-5,14,16-18,20; Samples 6-13,15,19	1.00	0.995
10	24.28	0.09	Samples 1,2,7-10,12,16,17,19,20; Samples 3-6,11,13-15,18	0.99	0.960
11	9.15	0.09	Samples 1,4,8,11-16,19,20; Samples 2,3,5-7,9,10,17,18	0.99	0.968
12	35.74	0.09	Samples 1-3,7,8,11-13,15,20; Samples 4-6,9,10,14,16-19	1.00	0.996

[a] The entire hit list, containing 26 hits, is shown in Table C.3. Hits shaded in green had matching sample index assignments and were correctly clustered into the two simulated classes. The concentration ratio ([Class A]/[Class B]) and *p*-value obtained from a *t*-test is also provided.

Similarly, Table 5.2 shows the top 12 peaks in the hit list for the chromatographic simulation shown in Figure 5.1B, where the $V_{R,B}$ for each analyte could vary between 0.09 and 0.25. The entire hit list for the 25 discovered peaks is shown in Table C.5. Here, it is not obvious which hits are the six class-indicating analytes based solely on the RSD^2 metric. Table 5.2 shows that Hits #1-4, 7, and 12 have matching sample index assignments at $k = 2$ (shaded in green) while none of the hits had matching index assignments at $k = 3$ (Table C.6). Applying Eqs. 5.3 and 5.4, the probability of six peaks having matching index assignments out of 25 total hits occurring by chance is 3.06×10^{-22}. Given this unlikely probability that these matches occurred by chance, it was assumed that meaningful sample-to-sample differences were causing the

clustering. The sample index assignments exhibited by Hits #1-4,7, and 12 were then applied to all peaks in the hit list (Table 5.2) to calculate a concentration ratio and *p*-value for each hit. Hits #1-4, 7, and 12 were all found to be statistically significant ($p < 0.05$) with a concentration ratio matching the changes simulated in the peak heights at those locations. Notably, analytes with either a 2-fold or 3-fold concentration change were still at the top of the hit list (Table 5.2). However, Table 5.2 shows that the peaks that changed by a factor of 1.5 were lower on the hit list (Hits #7 and #12) because the other peaks had a larger RSD^2 in comparison due to a randomly assigned large $V_{R,B}$ contribution. Therefore, it is important to mine the entire hit list with VRI-USI to find analytes that reveal the hidden groupings of the samples in the data.

Table 5.2. Resulting hit list after application of VRI-USI to the simulated data set containing a background variance of 0.09 to 0.25 (30 – 50 % *RSD*), which is shown in Figure 5.1B. The top 12 hits, ranked by RSD^2, are shown for brevity.[a]

Hit Number	t_R (s)	RSD^2	Sample Index Assignments	[Class A]/[Class B]	*p*-value
1	15.51	0.56	Samples 1-10; Samples 11-20	2.96	< 0.001
2	25.34	0.40	Samples 1-10; Samples 11-20	0.34	< 0.001
3	5.88	0.33	Samples 1-10; Samples 11-20	0.50	< 0.001
4	16.96	0.28	Samples 1-10; Samples 11-20	2.01	0.001
5	12.33	0.24	Samples 1-7,9,10,14; Samples 8,11-13,15-20	1.00	0.998
6	7.03	0.24	Samples 1,5,6,8,9,11,14,16-19; Samples 2-4,7,10,12,13,15,20	1.00	0.994
7	33.82	0.23	Samples 1-10; Samples 11-20	1.50	0.045
8	32.08	0.22	Samples 1-3,9-12,15,19; Samples 4-8,13,14,16-18,20	1.00	0.997
9	30.64	0.22	Samples 1-9,11,20; Samples 10,12-19	1.00	0.998
10	40.17	0.21	Samples 1-4,8,9,11-13,15; Samples 5-7,10,14,16-20	1.00	0.998
11	4.43	0.20	Samples 1,5,6,9-11,13,15,16; Samples 2-4,7,8,12,14,17-20	1.00	0.992
12	29.00	0.19	Samples 1-10; Samples 11-20	0.67	0.032

[a] The entire hit list, containing 25 hits, is shown in Table C.4. Hits shaded in green had matching sample index assignments and were correctly clustered into the two simulated classes. The concentration ratio ([Class A]/[Class B]) and *p*-value obtained from a *t*-test is also provided.

Since each of the simulated data sets had a random assignment of peak heights, mass spectra, and retention times, the retention times of the features discovered by VRI-USI were compared with the simulated retention times of the class-indicating features. For all 50 simulations containing a $V_{R,B}$ of 0.09, VRI-USI was able to identify an average of 5.4 peaks out of the original 6 class-indicating peaks (90 %). For the 50 simulations containing a variable $V_{R,B}$, VRI-USI was able to identify an average of 5.1 peaks out of the original 6 class-indicating peaks (85 %). It is important to note that analytes with a 1.5-fold concentration change were the least discoverable by VRI-USI in simulations containing a variable $V_{R,B}$ because the $V_{R,B}$ could have been larger than $V_{R,Chem}$. Regardless, the VRI-USI workflow was able to identify a majority of class-indicating analytes while operating in an unsupervised manner.

Often, the final goal of data analysis workflows is to develop chemometric models for sample differentiation using only the relevant features discovered [7,16]. PCA was performed on each simulated data set before and after application of VRI-USI, and *k*-means clustering was again used to determine if the samples clustered into the correct class labels. Figure 5.2A-B shows the scores plot using all the data for the respective simulations containing a constant and variable $V_{R,B}$, which are shown in Figure 5.1A-B. Prior to applying any feature selection, poor class separation was observed in the scores plots; when applying *k*-means clustering to the scores plots at $k = 2$, the samples do not correctly cluster into their two classes. Overall, *k*-means correctly clustered the samples into Class A and B for 11 out of 50 data sets (22 %) and 4 out of 50 data sets (8 %) for the respective constant and variable $V_{R,B}$ simulations when using all the data to build the PCA models. Clearly, the chemically insignificant within-class "background" variation masks the chemically significant variation for the six class-indicating analytes. After applying VRI-USI, only those *m/z* and timepoints that had matching sample index assignments

and were statistically significant ($p < 0.05$) were included for PCA. The resulting scores plots for both the constant and variable $V_{R,B}$ simulation (Figure 5.1A-B) are shown in Figure 5.2C-D. Here, *k*-means clustering was able to correctly group these samples into Class A and B. Impressively, for both the constant and variable $V_{R,B}$ simulations, *k*-means correctly clustered the samples into the two classes for all data sets (100 %) using the PCA models built with the features identified with VRI-USI. Since unsupervised tools like PCA will be more successful at differentiating two classes of GC-MS data when noisy and irrelevant variables are not included, these results serve to validate that VRI-USI is well suited to find the underlying chemical differences between the samples.

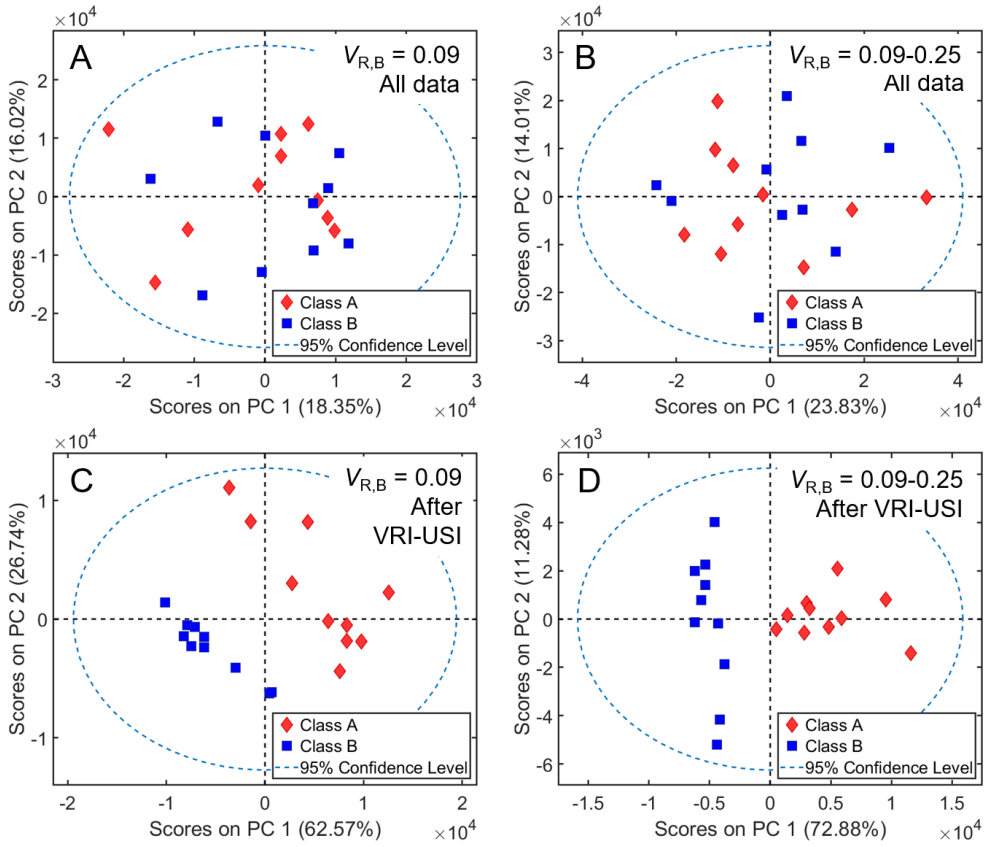


Figure 5.2. PCA scores plots for chromatographic simulations containing a background variance ($V_{R,B}$) of 0.09 (A,C) and a background variance ($V_{R,B}$) of 0.09 – 0.25 (B,D). The data set shown is from simulation #13, illustrated in Figure 5.1. (A,B) PCA scores plots using the data for all 50 peaks. (C,D) PCA scores plots using only the data for the six features that were discovered by VRI-USI. Class A and B are shown as red diamonds and blue squares, respectively.

5.4.2. Evaluation of VRI-USI with yeast metabolome data set

A previously studied data set of fermenting and respiring yeast [34–36] was used to study the application of VRI-USI to natural, complex samples. The overlaid TIC chromatograms of the 6 R and 6 DR yeast extract samples in the reduced GC-MS format are shown in Figure 5.3A and Figure 5.3B, respectively. This data set was chosen because the biological variation observed between yeast cultures was the primary contributor to the total variance in the data set, with an average *RSD* between 30 % and 50 % [34]. The identification of 54 class-distinguishing and 19 false positive metabolites has been well-documented in previous studies [35,36]. For example, the six metabolites labeled in Figure 5.3 (glycerol, threonine, malate, 5-oxoproline, glucose, and trehalose) represent class-distinguishing analytes with varied DR to R concentration ratios, [DR]/[R], between classes [35,36]. Previous non-targeted studies of this data set were able to identify these class-distinguishing metabolites even though the presence of false positive metabolites like ornithine and isoleucine were also present, making data mining efforts onerous [35,36].

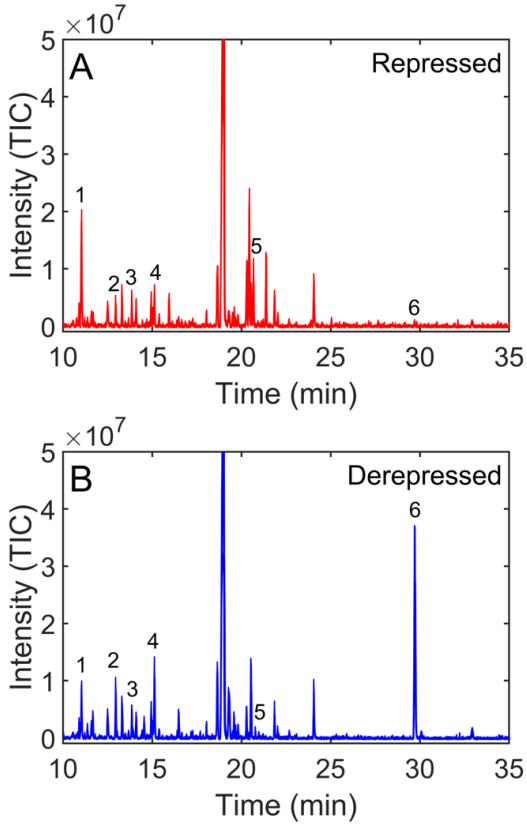


Figure 5.3. Overlaid 1D TIC chromatograms (10-35 min) of repressed (A; red) and derepressed (B; blue) yeast metabolome samples. Analytes of interest are numbered: (1) glycerol at 663 s, (2) threonine at 777 s, (3) malate at 873 s, (4) 5-oxoproline at 907.5 s, (5) glucose at 1240.5 s, and (6) trehalose at 1782.75 s.

To demonstrate the chromatographic variance seen for the six labeled analytes in Figure 5.3, the mean and standard deviation of the peak height per-*m/z* for each metabolite *within* each class were calculated. Figure 5.4A shows the scatter plot of these measurements for each analyte in two classes (see Figure C.1 and Figure C.2 for plots with each class and analyte displayed separately), where a general linear trend can be observed. To confirm this observation, the data were transformed logarithmically and graphed in a log-log plot (Figure 5.4B). Once transformed, the data appear linear, except for a region where the standard deviation levels off at lower peak heights. This region can be thought of as being "detection variation dominated" ($V_{R,Det}$), which is consistent with a significant contribution of $V_{R,Det}$ relative to the other sources of variation at low

signals approaching the *LOD*. Setting a signal threshold (black dashed line) in this region will reduce the presence of artificially high relative variances due to low mean signals, which could potentially be false positives. A line of best with a slope ≈ 1 was fit to the data above the signal threshold (red line and equation shown in Figure 5.4B), indicating that the standard deviation increases linearly with the signal. Hence, the *RSD* (i.e., the slope of a line defining the standard deviation versus the mean peak height) should be essentially constant for peak heights above the signal threshold. This is confirmed in Figure 5.4C, where a relatively constant band of points above the signal threshold is observed, followed by an increase in *RSD* below the threshold. The average *RSD* above the signal threshold is 0.22 (red solid line in Figure 5.4C) and the *RSD* above the signal threshold ranges from approximately 0.1 to 0.5, which corresponds to the range of biological variation previously observed [34]. Therefore, use of a signal threshold ensures that analytes near the top of the VRI-USI hit list have a $V_{R,Chem}$ contribution instead of a large $V_{R,Det}$ contribution.

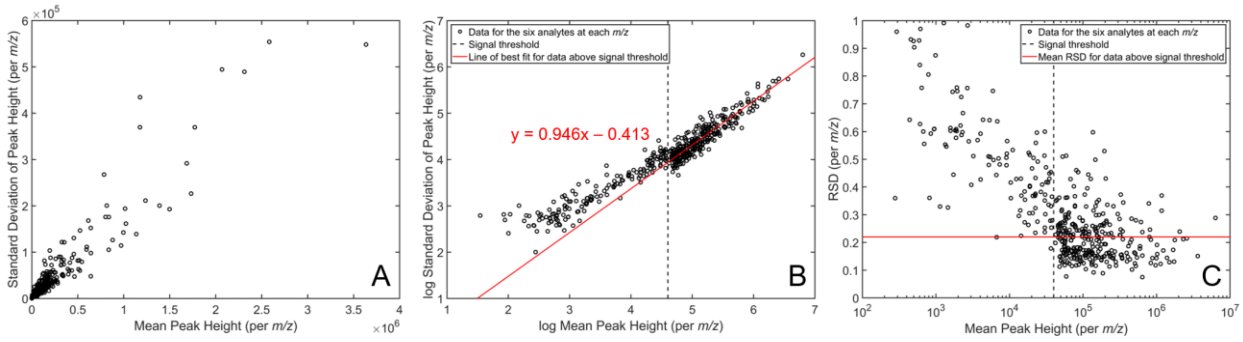


Figure 5.4. (A) Scatter plot of the standard deviation versus mean of the peak height for each *m/z* measured for the six analytes indicated in Figure 5.3 in the two different classes. Similar scatterplots broken down by class and analyte are in Figure C.1 and Figure C.2. One data point with a mean peak height of ~6.3×10^7 and standard deviation of ~1.8×10^6 was left out. (B) Logarithmically transformed standard deviation and mean peak height data from (A). The signal threshold applied to the data on a per *m/z* basis (~4×10^4) is shown (dotted line). A line of best fit (red solid line) was fitted through the data above the signal threshold and the equation is given. (C) *RSD* versus mean of the peak height for the *m/z* of the six analytes with *RSD* between 0 and 1. The average *RSD* of the data above the signal threshold (black dotted line) is 0.22 (red solid line).

Table 5.3 lists 53 peaks detected in the yeast metabolome data set, ranked by their average RSD^2. Probable identification of these peaks was accomplished by reviewing the metabolite identities at similar retention times previously reported [34-36] and was supported by comparisons of *m/z* included for each analyte with reference mass spectra. However, as seen in previous work [34-36], some of the peaks were unable to be identified with any level of certainty. The sample index assignments for the peaks in Table 5.3 were computed using *k*-means clustering at $k = 2$. Nineteen peaks out of 53 hits had matching sample index assignments (shaded in green), which the probability of these matches occurring by chance is 5.15×10^{-43}. Table 5.3 also shows that other peaks near the bottom of the hit list had matching sample index assignments (shaded in orange, yellow, or blue) that differ from the 19 peaks shaded in green. The probability of these other matching sample index assignments occurring by chance is 3.81×10^{-7} (orange) and 8.77×10^{-11} (yellow and blue). Given the substantially higher probability of the peaks shaded in orange, yellow, and blue, these additional matches are primarily due to random chance. Matching sample index assignments can also be observed at $k = 3$; however, many of these matches are also seen near the bottom of the hit list and/or have a comparably larger probability of occurring by chance (Table C.7). It is hypothesized that using the other index assignments with a high probability will cause the top hits to not be statistically significant even though their large RSD^2 indicates a large $V_{R,Chem}$ contribution. Therefore, the sample index assignments for the analytes shaded in green at $k = 2$ was assumed for calculating the concentration ratio and *t*-test results for all hits in Table 5.3.

Table 5.3. Results of VRI-USI, ranked by RSD^2, to the peak table for the yeast metabolome data set.[a]

Hit Number	Analyte	t_R (s)	RSD^2	Sample Index Assignments	[DR]/[R]	*p*-value
1	Unk1	956.25	1.83	Samples 1-6; Samples 7-12	R only	0.0122
2	Unk2	1220.625	1.63	Samples 1-6; Samples 7-12	R only	0.0065
3	Unk3	1503	1.47	Samples 1-6; Samples 7-12	R only	0.0011
4	Glucopyranose	1282.5	1.38	Samples 1-6; Samples 7-12	R only	0.0049
5	Glucose	1240.5	1.18	Samples 1-6; Samples 7-12	R only	0.0004
6	Glucose	1227	1.17	Samples 1-6; Samples 7-12	R only	0.0001
7	Trehalose	1782.75	1.11	Samples 1-6; Samples 7-12	68.2	< 0.0001
8	Unk4	699.75	0.836	Samples 1-3,5-10; Samples 4,11,12	1.48	0.3231
9	Malate	873	0.809	Samples 1-6; Samples 7-12	9.49	0.0002
10	5'-S-Methyl-5'-thioadenosine	1804.5	0.799	Samples 1-6; Samples 7-12	6.16	0.0009
11	Unk5	924	0.687	Samples 1,3,6,7,9,11; Samples 2,4,5,8,10,12	0.735	0.5041
12	Homoserine	835.5	0.612	Samples 1-6; Samples 7-12	4.20	0.0006
13	Citrate	1160.25	0.562	Samples 1-6; Samples 7-12	3.51	0.0006
14	Tyrosine	1247.25	0.537	Samples 1-6; Samples 7-12	2.81	0.0005
15	Ornithine	1156.5	0.495	Samples 1-6,9; Samples 7,8,10-12	3.51	0.0703
16	Lysine	1232.25	0.439	Samples 1-6; Samples 7-12	1.86	0.0012
17	Unk6	1170.75	0.406	Samples 1-6; Samples 7-12	0.476	0.0016
18	Unk7	696	0.404	Samples 1,3,6,7,9; Samples 2,4,5,8,10-12	1.21	0.3016
19	Glucopyranose	1216.5	0.404	Samples 1-6; Samples 7-12	R only	0.0033
20	Unk8	1323	0.384	Samples 1,3,6,7,9,12; Samples 2,4,5,8,10,11	0.695	0.1896
21	Unk9	831	0.355	Samples 3-5,10,11; Samples 1,2,6-9,12	1.16	0.6960
22	Glutamic acid	988.5	0.349	Samples 1-6; Samples 7-12	3.12	0.0001
23	Unk10	1175.25	0.324	Samples 1-6; Samples 7-12	2.24	0.0019
24	Glycerol	663	0.303	Samples 1-6; Samples 7-12	0.413	< 0.0001
25	Unk11	648	0.268	Samples 1,3,5-7,9,11; Samples 2,4,8,10,12	1.16	0.6403
26	Threonine	777	0.247	Samples 1-6; Samples 7-12	2.55	0.0001
27	5-Oxoproline	907.5	0.237	Samples 1-7,9; Samples 8,10-12	2.35	0.0007
28	Unk12	1082.25	0.153	Samples 1,3,5-7,9,11; Samples 2,4,8,10,12	1.18	0.4098
29	Unk13	984	0.136	Samples 1-6,12; Samples 7-11	2.05	0.2091
30	Asparginine	1036.5	0.135	Samples 1,3,6,7,9-11; Samples 2,4,5,8,12	0.833	0.3156
31	Unk14	781.5	0.120	Samples 1,2,4-6; Samples 3,7-12	2.01	0.0001
32	Unk15	1030.5	0.112	Samples 1,3,5,6,9,10; Samples 2,4,7,8,11,12	1.19	0.2393
33	Isoleucine	682.5	0.111	Samples 1,2,5,6,9; Samples 3,4,7,8,10-12	1.59	0.1061
34	Unk16	1974.75	0.106	Samples 1-3,6,7,9,11; Samples 4,5,8,10,12	1.53	0.0082
35	Unk17	902.25	0.100	Samples 1,3,6,7,9,11,12; Samples 2,4,5,8,10	0.973	0.7725
36	Unk18	1360.5	0.0918	Samples 1,3,6,7,9,11; Samples 2,4,5,8,10,12	1.05	0.4617
37	Methionine	897	0.0792	Samples 1,3,5-7,9,11; Samples 2,4,8,10,12	1.09	0.5529
38	Unk19	798.75	0.0722	Samples 1,3-7,9; Samples 2,8,10-12	1.24	0.1058
39	Leucine	654	0.0708	Samples 1,3,6,7,11; Samples 2,4,5,8-10,12	0.748	0.0738
40	Phenylalanine	999	0.0642	Samples 1,5-7,9,10,12; Samples 2-4,8,11	1.03	0.8640
41	Unk20	1188.75	0.064	Samples 1,3,5-7,9,11; Samples 2,4,8,10,12	0.826	0.2155
42	Unk21	1140	0.0575	Samples 1,3,6,7,9-11; Samples 2,4,5,8,12	0.951	0.5993
43	Unk22	1135.5	0.0558	Samples 1,3,5-7,9,12; Samples 2,4,8,10,12	0.954	0.7090
44	Unk23	820.5	0.0548	Samples 1,3,6,7,9-11; Samples 2,4,5,8,12	0.909	0.4584
45	Glutamine	1119	0.0542	Samples 1,3,5-7,9,11; Samples 2,4,8,10,12	1.21	0.1503
46	Unk24	810	0.0520	Samples 1,3,6,7,9-11; Samples 2,4,5,8,12	0.916	0.5070
47	Unk25	633	0.0488	Samples 1,3,5-7,9; Samples 2,4,8,10-12	1.06	0.5182

48	o-Toluic acid	750	0.0468	Samples 1,5-7,9; Samples 2-4,8,10-12	1.03	0.7798
49	Unk26	846	0.0448	Samples 1,3,6,7,9-11; Samples 2,4,5,8,12	0.875	0.2502
50	Unk27	673.5	0.0435	Samples 1,3,6,7,9-11; Samples 2,4,5,8,12	0.871	0.1284
51	Unk28	1136.25	0.0351	Samples 1,3,6,7,9,11; Samples 2,4,5,8,10,12	0.955	0.6540
52	Unk29	1311	0.0312	Samples 1,3,6,7,9,11; Samples 2,4,5,8,10,12	0.916	0.3773
53	Stearic acid	1443	0.0299	Samples 1,3,5-7,9,11; Samples 2,4,8,10,12	1.12	0.2083

[a] Hits shaded in green had matching sample index assignments and were correctly clustered into the DR and R classes. Hits shaded in orange, yellow, and blue were not correctly clustered DR and R classes but did have matching sample index assignments. The concentration ratio ([DR]/[R]) and *p*-value obtained from a *t*-test is also provided.

Using the sample index assignments given by the analytes shaded in green, concentration ratios and *p*-values from a *t*-test were calculated (Table 5.3). Beneficially for analytes that could be identified after the reduction in dimensionality, their concentration ratios are approximately equal to the previously measured ratio between the DR and R classes, [DR]/[R] [35,36]. This result indicates that the sample index assignments with the lowest probability of occurring by chance highlighted the class-distinguishing differences in the data set. A *t*-test found that 22 of those 53 designated features were statistically significant ($p < 0.05$) using the assumed index assignments from *k*-means clustering. Of the 22 statistically significant peaks, 15 could be confidently identified (Table 5.3) and were all discovered to be class-distinguishing in previous studies [35,36]. Likewise, eight analytes that were confidently identified and had *p*-values less than 0.05 (Table 5.3) were previously identified to be false positives [35,36]. Figure 5.5 shows the relationship between RSD^2 and *p*-value for all 53 peaks, where a *p*-value of 0.05 is indicated by a dot-dashed blue line. The 19 peaks that had matching sample index assignments (shaded in green) are represented by a filled circle in Figure 5.5, while peaks shown by an unfilled circle did not have those index assignments after *k*-means clustering. Generally, the yeast metabolome data set shows that peaks with a large RSD^2 were likely to have a *p*-values < 0.05 and matching sample index assignments. However, Figure 5.5 illustrates the three peaks (Hits #27, 31, and 34) that were found to be statistically significant ($p < 0.05$) did not have the same sample index

assignments as the other 19 statistically significant peaks, indicating that the natural groupings discovered by *k*-means clustering differ from the experimental labels.

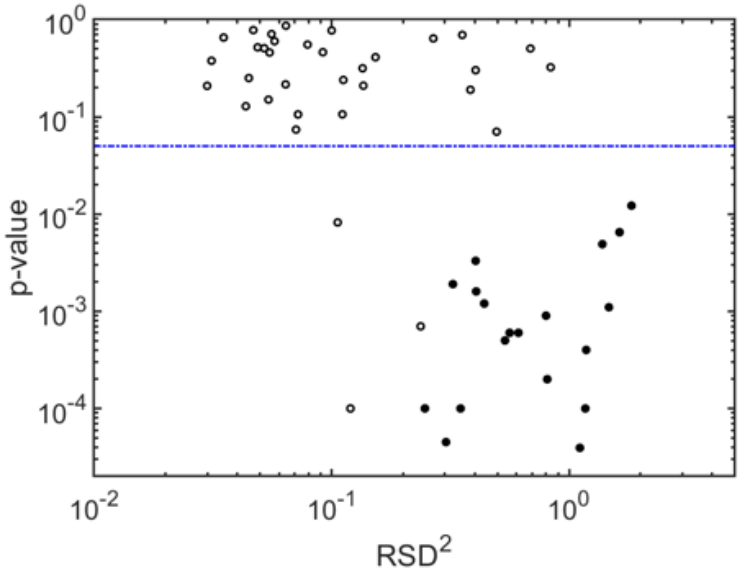


Figure 5.5. Scatter plot of *p*-value versus RSD^2 for the 53 peaks identified in the yeast metabolome data set. Filled circles represent the 19 peaks with matching sample index assignments (shaded in green in Table 5.3) after *k*-means clustering while unfilled circles represent the other 34 peaks. The dot-dashed blue line represents a *p*-value of 0.05.

Figure 5.6 shows the overlaid analytical ion current (AIC) chromatograms of the R (red) and DR (blue) classes for four representative metabolites: malate (A), glycerol (B), threonine (C), and 5-oxoproline (D). The AICs were obtained by summing the signal for the *m/z* that defined each peak. A variety of concentration ratios, [DR]/[R], can be seen for the four analytes shown here in Figure 5.6 and for all statistically significant peaks in Table 5.3. Analytes that were only present in one class or had large [DR]/[R] were found to have a large RSD^2. For example, malate (Figure 5.6A), had prominent signal in the DR class with a concentration ratio of 9.49, and had a RSD^2 of 0.809. Meanwhile, Figure 5.6 shows that glycerol (B), threonine (C), and 5-oxoproline (D) had smaller differences between the DR and R classes, so these analytes had a small RSD^2. All the peaks shown in Figure 5.6 had matching sample index assignments

after *k*-means clustering except for 5-oxoproline. The overlaid AIC chromatogram for 5-oxoproline (Figure 5.6D) shows that two samples in the DR class are of similar signal to the samples in the R class, which caused the unsupervised *k*-means algorithm to cluster these two DR samples with the R class (Table 5.3). However, after assuming the sample index assignments with the lowest probability, 5-oxoproline was determined to be statistically significant ($p < 0.05$). This result illustrates that using the VRI-USI workflow can provide insight into both the underlying data structure for individual peaks and chemical differences among samples.

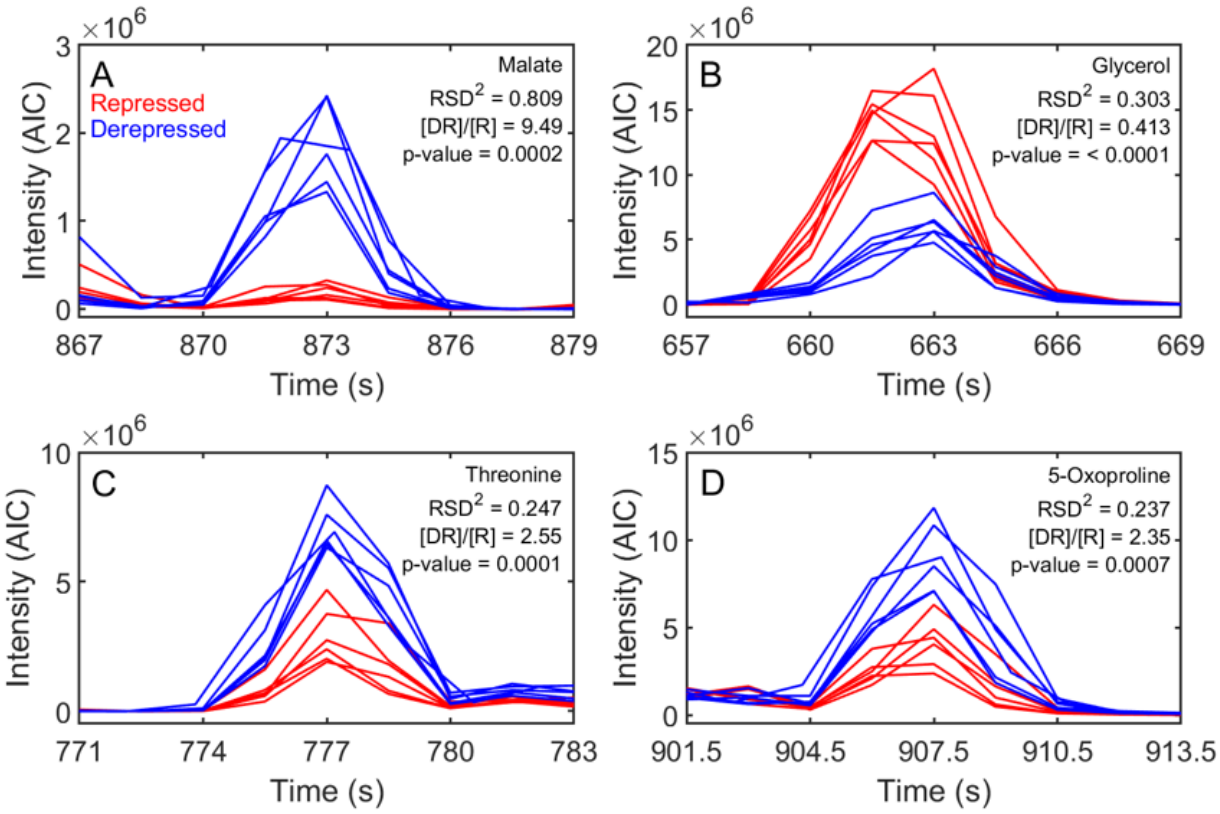


Figure 5.6. Analytical ion current (AIC) chromatograms of the repressed (red) and derepressed (blue) classes for four identified metabolites: (A) malate, (B) glycerol, (C) threonine, and (D) 5-oxoproline. The measured RSD^2, concentration ratio, and *p*-values are also provided.

5.4.3. Evaluation of VRI-USI with human cancer data set

Application of unsupervised workflows to human metabolomics studies has been beneficial for exploratory data analysis without introducing human biases [17,19]. However,

identification of discriminatory analytes without the use of class labels can be hampered due to inter-individual differences [19,46,47]. Herein, VRI-USI was applied to a previously collected data set that investigated the differences in volatile salivatory analytes based on the presence of head and neck cancer [37]. Human saliva of 32 cancer and 27 control patients was analyzed using HS-SPME-GC-MS [37]. Figure 5.7 compares the TIC chromatograms for control patient #7 (A; red) and cancer patient #6 (B; blue). Of the 48 analytes identified in the study, 27 were discovered to be statistically significant ($p < 0.05$) [37]. For example, the five analytes labeled in Figure 5.7 (ethyl propanoate, 1,4-dichlorobenzene, acetic acid, 1,2-decanediol, and 2,5-di-*tert*-butylphenol) were statistically different between the classes.

The resulting hit list after application of VRI-USI to the 48 peaks identified in this human metabolomics data set is shown in Table 5.4. When comparing the sample index assignments generated at $k = 2$, five metabolites (shaded in green) were found to have matching sample index assignments (Table C.8). These five metabolites (ethyl propanoate, 1,4-dichlorobenzene, acetic acid, 1,2-decanediol, and 2,5-di-*tert*-butylphenol) are also labeled in Figure 5.7. Interestingly, these metabolites are not concentrated near the very top of the hit list like the previous data sets; instead, they are intermingled throughout the hit list with their RSD^2 ranging from 1.21 to 3.87. The sample index assignments for the top hits in the hit list (Table C.8) illustrate the large human-to-human variability and possibility of outliers seen in human metabolomics studies [19,46,47]. While these five matching metabolites were not near the very top of the hit list, the probability of these matches occurring by chance is 6.45×10^{-78}. While only a small number of metabolites with matching sample index assignments were found after applying VRI-USI to this data set, it is important to note that these matches are due to chemical differences between samples and not random correlations because of the large number of samples in the data set.

Table C.9 also illustrates that none of the analytes have matching index assignments at $k = 3$. Using the sample index assignments of the five matching analytes (Table 5.4), a wide range of concentration ratios, ranging from 0.05 to 32.1, was observed. Additionally, 25 of the benchmarked 27 class-distinguishing metabolites [37] were discovered herein.

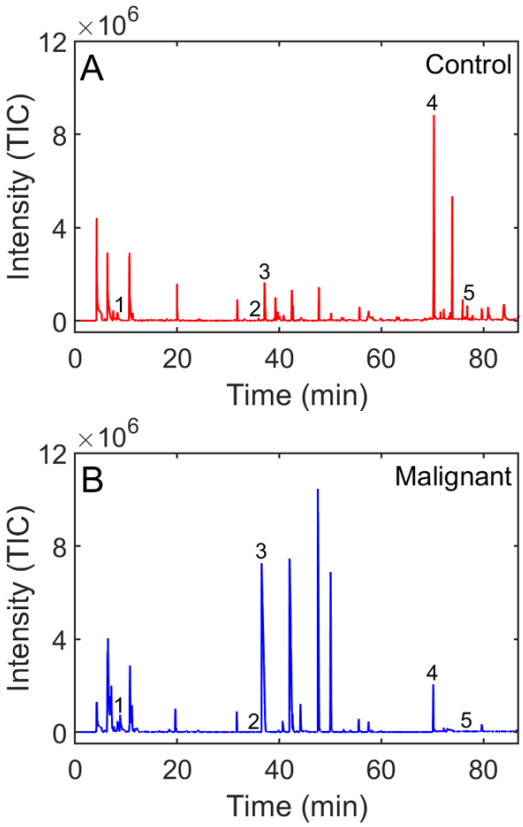


Figure 5.7. Representative TIC chromatograms of salivary profile of control patient #7 (A) and a head and neck cancer patient #6 (B). Analytes of interest are numbered: (1) ethyl propanoate at 8.7 min, (2) 1,4-dichlorobenzene at 36 min, (3) acetic acid at 37.1 min, (4) 1,2-decanediol at 70.3 min, and (5) 2,5-di-*tert*-butylphenol at 76.8 min.

Table 5.4. Results of VRI-USI, ranked by RSD^2, to the peak table for the head and neck cancer data set.[a]

Hit Number	Analyte Name	RSD^2	[Malignant]/[Control]	*p*-value
1	1-Propanol	10.62	5.61	0.101
2	Acetone	10.22	2.96	0.227
3	*p*-Cresol	7.37	10.7	0.021
4	*o*-Cymene	6.63	0.97	0.961
5	Pentanoic acid, 4-methyl-	6.39	32.1	0.005
6	*p*-Cymene	6.16	0.94	0.928
7	2-Decanone	5.55	4.65	0.035
8	Propanoic acid, 2-methyl-	4.40	29.3	0.001
9	Phenol	4.33	4.92	0.014
10	*o*-Xylene	3.95	3.52	0.030
11	Phenol, 2,5-bis(1,1-dimethylethyl)-	3.87	0.08	0.001
12	Acetic acid ethenyl ester	3.23	4.55	0.006
13	Propanoic acid, ethyl ester	3.02	11.1	< 0.001
14	Octane, 3,3-dimethyl-	2.89	2.10	0.100
15	*n*-Propyl acetate	2.86	4.96	0.003
16	2,4-Dimethyl-1-heptene	2.65	3.48	0.009
17	Ethyl Acetate	2.43	7.45	< 0.001
18	Acetic acid	2.36	17.7	< 0.001
19	2,3-Pentanedione	2.33	1.41	0.410
20	Benzaldehyde	2.30	1.61	0.221
21	Ethylbenzene	2.24	1.48	0.305
22	1-Butanol	2.06	2.35	0.028
23	Benzaldehyde, 3-methyl-	1.76	2.50	0.012
24	Benzene, 4-ethenyl-1,2-dimethyl-	1.74	0.64	0.190
25	Propanoic acid	1.74	2.16	0.410
26	Ethanol	1.61	0.64	0.198
27	Styrene	1.60	2.44	0.010
28	Furfural	1.53	0.81	0.528
29	3-Furaldehyde	1.53	0.83	0.568
30	2-Propanol, 1-chloro-	1.48	2.44	0.007
31	1,2-Decanediol	1.45	0.05	< 0.001
32	Decane, 4-methyl-	1.32	1.14	0.662
33	*p*-Xylene	1.28	0.89	0.678
34	Pentane, 2,3,3-trimethyl-	1.22	3.42	< 0.001
35	Benzene, 1,4-dichloro-	1.21	0.11	< 0.001
36	Hexane, 3-methyl-	0.83	2.12	0.002
37	3-Pentenoic acid, 4-methyl-	0.83	1.27	0.316
38	1-Dodecanol, 3,7,11-trimethyl-	0.82	1.50	0.089
39	2-Butanol, 1-chloro-	0.78	1.73	0.017
40	Propanoic acid, propyl ester	0.78	1.00	0.992
41	Butane, 1,2,3,4-tetrachloro-	0.77	0.95	0.809
42	3-Decen-2-ol, (*E*)-	0.72	0.84	0.418
43	Benzene, 1,3-bis(1,1-dimethylethyl)-	0.60	0.80	0.248

44	Propane, 1,2,3-trichloro-2-methyl-	0.48	1.23	0.255
45	Propane, 1,1,3,3-tetrachloro-2-methyl-	0.46	0.51	< 0.001
46	Toluene	0.40	1.04	0.786
47	2-Propanol, 1,3-dichloro-	0.36	0.67	0.011
48	2-Propenoic acid	0.29	0.57	< 0.001

[a] The sample index assignments are shown in Table C.5. Hits shaded in green had matching sample index assignments and were correctly clustered into the malignant and control classes. The concentration ratio ([Malignant]/[Control]) and *p*-value obtained from a *t*-test is also provided.

Figure 5.8A compares the PCA scores plot prepared using the signals from all the metabolites listed in Table 5.4 versus the PCA scores plot in Figure 5.8B using just the five metabolites that had matching sample index assignments at $k = 2$ (ethyl propanoate, 1,4-dichlorobenzene, acetic acid, 1,2-decanediol, and 2,5-di-*tert*-butylphenol). Here, the samples are labeled according to the two classes, where malignant and control samples are shown as red diamonds and blue squares, respectively. Figure 5.8A illustrates that inter-individual differences in the cancer patient samples causes poor class separation, while Figure 5.8B shows a high degree of class separation between the malignant and control samples, due to the significant predictive power of the five metabolites. Utilizing the sample index assignments for these metabolites, *k*-means clustering separates the 32 malignant samples from the 27 control samples (Table C.8). Previous work has shown that the performance of *k*-means clustering worsens when data sets have unequal class sizes; however, the use of different methods for selecting the initial centroid locations and/or performing multiple iterations of *k*-means clustering can improve performance on unbalanced data sets [48,49]. Herein, multiple iterations of *k*-means clustering were performed since initial centroid locations were chosen at random. Therefore, the VRI-USI workflow can be adapted to unbalanced data sets to discern class-based differences in an unsupervised fashion.

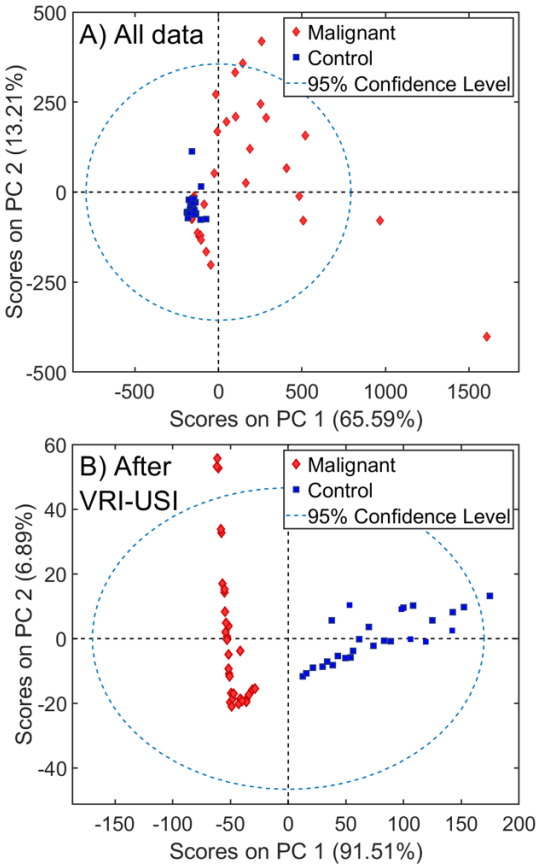


Figure 5.8. PCA score plot prepared using all 48 metabolites identified in Table 5.4. (B) PCA scores plot prepared using the five analytes that had matching sample index assignments discovered by VRI-USI (ethyl propanoate, 1,4-dichlorobenzene, acetic acid, 1,2-decanediol, and 2,5-di-*tert*-butylphenol). Malignant and control samples are shown as red diamonds and blue squares, respectively.

5.5. Conclusion

An unsupervised data analysis method, referred to as VRI-USI, was proposed and demonstrated for GC-MS data. The first step of this method discovers peaks with large signal variances for a given analyte peak between the samples, which could potentially indicate a chemical difference. Using the peak profiles developed for these discovered peaks, *k*-means clustering is utilized to group the samples into different clusters for each analyte. We illustrate that the sample index assignment that has the lowest probability of occurring by chance is most likely to indicate the chemical differences between samples, revealing the true structure of the

data set. GC-MS data sets with significant within-class background variation were simulated and VRI-USI was applied in a pixel-based fashion. VRI-USI discovered approximately 85 – 90 % of the peaks designated to change between classes. Using a peak table-based approach, VRI-USI was also applied to a previously studied yeast metabolome data set [34-36] to further validate this unsupervised chemometric workflow. Ultimately, 22 peaks were found to be statistically different ($p < 0.05$) using the sample cluster indices with the lowest probability of occurring due to random correlations. Similarly, when VRI-USI was applied to a peak-table developed for a human cancer metabolomics study [37], 25 peaks were identified to be statistically different ($p < 0.05$). The overall sample index assignments for each data set were validated to be correct using knowledge of class membership. Note, the data sets in this proof-of-concept study naturally had two classes. For data sets that naturally have more classes, we hypothesize that peaks near the top of the hit list will have the largest sample-based differences; however, *k*-means clustering with more clusters will still reveal the natural data structure. Future work will explore the application of VRI-USI to data sets with more classes and to GC×GC-TOFMS data sets [50,51].

5.6. References

[1] F.J. Santos, M.T. Galceran, Modern developments in gas chromatography-mass spectrometry-based environmental analysis, J. Chromatogr. A 1000 (2003) 125–151. https://doi.org/10.1016/S0021-9673(03)00305-4.

[2] M.M. Koek, R.H. Jellema, J. van der Greef, A.C. Tas, T. Hankemeier, Quantitative metabolomics based on gas chromatography mass spectrometry: Status and perspectives, Metabolomics 7 (2011) 307–328. https://doi.org/10.1007/s11306-010-0254-3.

[3] A. Chauhan, M.K. Goyal, P. Chauhan, GC-MS Technique and its Analytical Applications in Science and Technology, J. Anal. Bioanal. Tech. 5 (2014). https://doi.org/10.4172/2155-9872.1000222.

[4] H. Song, J. Liu, GC-O-MS technique and its applications in food flavor analysis, Food Res. Int. 114 (2018) 187–198. https://doi.org/10.1016/j.foodres.2018.07.037.

[5] K.M. Pierce, J.L. Hope, K.J. Johnson, B.W. Wright, R.E. Synovec, Classification of gasoline data obtained by gas chromatography using a piecewise alignment algorithm combined with feature selection and principal component analysis, J. Chromatogr. A 1096 (2005) 101–110. https://doi.org/10.1016/j.chroma.2005.04.078.

[6] N.E. Watson, M.M. VanWingerden, K.M. Pierce, B.W. Wright, R.E. Synovec, Classification of high-speed gas chromatography-mass spectrometry data by principal component analysis coupled with piecewise alignment and feature selection, J. Chromatogr. A 1129 (2006) 111–118. https://doi.org/10.1016/j.chroma.2006.06.087.

[7] L.A. Adutwum, J.J. Harynuk, Unique Ion Filter: A Data Reduction Tool for GC/MS Data Preprocessing Prior to Chemometric Analysis, Anal. Chem. 86 (2014) 7726–7733. https://doi.org/10.1021/ac501660a.

[8] C.E. Freye, P.R. Bowden, M.T. Greenfield, B.C. Tappan, Non-targeted discovery-based analysis for gas chromatography with mass spectrometry: A comparison of peak table, tile, and pixel-based Fisher ratio analysis, Talanta 211 (2020) 120668. https://doi.org/10.1016/j.talanta.2019.120668.

[9] C.N. Cain, N.J. Haughn, H.J. Purcell, L.C. Marney, R.E. Synovec, C.T. Thoumsin, S.C. Jackels, K.J. Skogerboe, Analytical Determination of the Severity of Potato Taste Defect in Roasted East African Arabica Coffee, J. Agric. Food Chem. (2021). https://doi.org/10.1021/acs.jafc.1c00605.

[10] C. Pizarro, I. Esteban-Díez, C. Sáenz-González, J.M. González-Sáiz, Vinegar classification based on feature extraction and selection from headspace solid-phase microextraction/gas chromatography volatile analyses: A feasibility study, Anal. Chim. Acta 608 (2008) 38–47. https://doi.org/10.1016/j.aca.2007.12.006.

[11] E. Kondo, P.J. Marriott, R.M. Parker, K.A. Kouremenos, P. Morrison, M. Adams, Metabolic profiling of yeast culture using gas chromatography coupled with orthogonal acceleration accurate mass time-of-flight mass spectrometry: Application to biomarker discovery, Anal. Chim. Acta 807 (2014) 135–142. https://doi.org/10.1016/j.aca.2013.11.004.

[12] L. Lebanov, L. Tedone, A. Ghiasvand, B. Paull, Random Forests machine learning applied to gas chromatography – Mass spectrometry derived average mass spectrum data sets for classification and characterisation of essential oils, Talanta 208 (2020) 120471. https://doi.org/10.1016/j.talanta.2019.120471.

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

[14] J.S. Ribeiro, F. Augusto, T.J.G. Salva, R.A. Thomaziello, M.M.C. Ferreira, Prediction of sensory properties of Brazilian Arabica roasted coffees by headspace solid phase microextraction-gas chromatography and partial least squares, Anal. Chim. Acta 634 (2009) 172–179. https://doi.org/10.1016/j.aca.2008.12.028.

[15] K.M. Pierce, S.P. Schale, Predicting percent composition of blends of biodiesel and conventional diesel using gas chromatography-mass spectrometry, comprehensive two-dimensional gas chromatography-mass spectrometry, and partial least squares analysis, Talanta 83 (2011) 1254–1259. https://doi.org/10.1016/j.talanta.2010.07.084.

[16] D.W. Cook, S.C. Rutan, Chemometrics for the analysis of chromatographic data in metabolomics investigations, J. Chemom. 28 (2014) 681–687. https://doi.org/10.1002/cem.2624.

[17] S. Ren, A.A. Hinzman, E.L. Kang, R.D. Szczesniak, L.J. Lu, Computational and statistical analysis of metabolomics data, Metabolomics 11 (2015) 1492–1513. https://doi.org/10.1007/s11306-015-0823-6.

[18] J. Li, K. Cheng, S. Wang, F. Morstatter, R.P. Trevino, J. Tang, H. Liu, Feature selection: A data perspective, ACM Comput. Surv. 50 (2017). https://doi.org/10.1145/3136625.

[19] J. Heinemann, Machine Learning in Untargeted Metabolomics Experiments, in: E.E.K. Baidoo (Ed.), Microb. Metabolomics Methods Protoc., Springer New York, New York, NY, 2019: pp. 287–299. https://doi.org/10.1007/978-1-4939-8757-3_17.

[20] A.K. Jain, Data clustering: 50 years beyond K-means, Pattern Recognit. Lett. 31 (2010) 651–666. https://doi.org/10.1016/j.patrec.2009.09.011.

[21] S. Wold, K. Esbensen, P. Geladi, Principal component analysis, Chemom. Intell. Lab. Syst. 2 (1987) 37–52. https://doi.org/10.1016/0169-7439(87)80084-9.

[22] P.E. Sudol, K.M. Pierce, S.E. Prebihalo, K.J. Skogerboe, B.W. Wright, R.E. Synovec, Development of gas chromatographic pattern recognition and classification tools for compliance and forensic analyses of fuels: A review, Anal. Chim. Acta 1132 (2020) 157–186. https://doi.org/10.1016/j.aca.2020.07.027.

[23] N. Pasadakis, E. Gidarakos, G. Kanellopoulou, N. Spanoudakis, Identifying sources of oil spills in a refinery by gas chromatography and chemometrics: A case study, Environ. Forensics 9 (2008) 33–39. https://doi.org/10.1080/15275920701729548.

[24] G. Aliakbarzadeh, H. Sereshti, H. Parastar, Pattern recognition analysis of chromatographic fingerprints of Crocus sativus L. secondary metabolites towards source identification and quality control, Anal. Bioanal. Chem. 408 (2016) 3295–3307. https://doi.org/10.1007/s00216-016-9400-8.

[25] S. Han, W. Zhang, P. Li, X. Li, J. Liu, B. Xu, D. Luo, Characterization of Aromatic Liquor by Gas Chromatography and Principal Component Analysis, Anal. Lett. 50 (2017) 777–786. https://doi.org/10.1080/00032719.2016.1196365.

[26] S.K. Jha, K. Hayashi, Molecular structural discrimination of chemical compounds in body odor using their GC–MS chromatogram and clustering methods, Int. J. Mass Spectrom. 423 (2017) 1–14. https://doi.org/10.1016/j.ijms.2017.09.010.

[27] M. Peikari, S. Salama, S. Nofech-Mozes, A.L. Martel, A Cluster-then-label Semi-supervised Learning Approach for Pathology Image Classification, Sci. Rep. 8 (2018) 1–13. https://doi.org/10.1038/s41598-018-24876-0.

[28] B.C. Kwon, B. Eysenbach, J. Verma, K. Ng, C. De Filippi, W.F. Stewart, A. Perer, Clustervision: Visual Supervision of Unsupervised Clustering, IEEE Trans. Vis. Comput. Graph. 24 (2018) 142–151. https://doi.org/10.1109/TVCG.2017.2745085.

[29] L. Haar, K. Anding, K. Trambitckii, G. Notni, Comparison between supervised and unsupervised feature selection methods, ICPRAM 2019 - Proc. 8th Int. Conf. Pattern Recognit. Appl. Methods. (2019) 582–589. https://doi.org/10.5220/0007385305820589.

[30] J. Brownlee, Data Preparation for Machine Learning: Data Cleaning, Feature Selection, and Data Transforms in Python, Machine Learning Mastery, 2020.

[31] X. He, D. Cai, P. Niyogi, Laplacian Score for Feature Selection, in: Y. Weiss, B. Schölkopf, J.C. Platt (Eds.), Adv. Neural Inf. Process. Syst. 18, MIT Press, 2006: pp. 507–514.

[32] D. Cai, C. Zhang, X. He, Unsupervised feature selection for Multi-Cluster data, Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min. (2010) 333–342. https://doi.org/10.1145/1835804.1835848.

[33] C. Ding, X. He, K-means clustering via principal component analysis, Proceedings, Twenty-First Int. Conf. Mach. Learn. ICML 2004. (2004) 225–232. https://doi.org/10.1145/1015330.1015408.

[34] R.E. Mohler, K.M. Dombek, J.C. Hoggard, E.T. Young, R.E. Synovec, Comprehensive two-dimensional gas chromatography time-of-flight mass spectrometry analysis of metabolites in fermenting and respiring yeast cells, Anal. Chem. 78 (2006) 2700–2709. https://doi.org/10.1021/ac052106o.

[35] R.E. Mohler, K.M. Dombek, J.C. Hoggard, K.M. Pierce, E.T. Young, R.E. Synovec, Comprehensive analysis of yeast metabolite GC×GC-TOFMS data: Combining discovery-mode and deconvolution chemometric software, Analyst 132 (2007) 756–767. https://doi.org/10.1039/b700061h.

[36] N.E. Watson, B.A. Parsons, R.E. Synovec, Performance evaluation of tile-based Fisher Ratio analysis using a benchmark yeast metabolome data set, J. Chromatogr. A 1459 (2016) 101–111. https://doi.org/10.1016/j.chroma.2016.06.067.

[37] R. Taware, K. Taunk, J.A.M. Pereira, A. Shirolkar, D. Soneji, J.S. Câmara, H.A. Nagarajaram, S. Rapole, Volatilomic insight of head and neck cancer via the effects observed on saliva metabolites, Sci. Rep. 8 (2018) 17725. https://doi.org/10.1038/s41598-018-35854-x.

[38] E. Boyaci, Á. Rodríguez-Lafuente, K. Gorynski, F. Mirnaghi, É.A. Souza-Silva, D. Hein, J. Pawliszyn, Sample preparation with solid phase microextraction and exhaustive extraction approaches: Comparison for challenging cases, Anal. Chim. Acta 873 (2015) 14–30. https://doi.org/10.1016/j.aca.2014.12.051.

[39] M. Lashgari, V. Singh, J. Pawliszyn, A critical review on regulatory sample preparation methods: Validating solid-phase microextraction techniques, TrAC - Trends Anal. Chem. 119 (2019) 115618. https://doi.org/10.1016/j.trac.2019.07.029.

[40] V.J. Barwick, Sources of uncertainty in gas chromatography and high-performance liquid chromatography, J. Chromatogr. A 849 (1999) 13–33. https://doi.org/10.1016/S0021-9673(99)00537-3.

[41] V.K. Rohatgi, A.K.M. Ehsanes Saleh, An Introduction to Probability and Statistics, Third Edit, John Wiley & Sons, Hoboken, NJ, 2015.

[42] F. Dondi, A. Bassi, A. Cavazzini, M.C. Pietrogrande, A Quantitative Theory of the Statistical Degree of Peak Overlapping in Chromatography, Anal. Chem. 70 (1998) 766–773. https://doi.org/10.1021/ac9705430.

[43] C.N. Cain, S. Schöneich, R.E. Synovec, Development of an Enhanced Total Ion Current Chromatogram Algorithm to Improve Untargeted Peak Detection, Anal. Chem. 92 (2020) 11365–11373. https://doi.org/10.1021/acs.analchem.0c02136.

[44] K. Haug, R.M. Salek, P. Conesa, J. Hastings, P. De Matos, M. Rijnbeek, T. Mahendraker, M. Williams, S. Neumann, P. Rocca-Serra, E. Maguire, A. González-Beltrán, S.A. Sansone, J.L. Griffin, C. Steinbeck, MetaboLights - An open-access general-purpose repository for metabolomics studies and associated meta-data, Nucleic Acids Res. 41 (2013) 781–786. https://doi.org/10.1093/nar/gks1004.

[45] R. Loohach, K. Garg, Effect of Distance Functions on Simple K-means Clustering Algorithm, Int. J. Comput. Appl. 49 (2012) 7–9. https://doi.org/10.5120/7629-0698.

[46] L.G. Rasmussen, F. Savorani, T.M. Larsen, L.O. Dragsted, A. Astrup, S.B. Engelsen, Standardization of factors that influence human urine metabolomics, Metabolomics 7 (2011) 71–83. https://doi.org/10.1007/s11306-010-0234-7.

[47] S.E. Prebihalo, G.S. Ochoa, K.L. Berrier, K.J. Skogerboe, K.L. Cameron, J.R. Trump, S.J. Svoboda, J.K. Wickiser, R.E. Synovec, Control-Normalized Fisher Ratio Analysis of Comprehensive Two-Dimensional Gas Chromatography Time-of-Flight Mass Spectrometry Data for Enhanced Biomarker Discovery in a Metabolomic Study of Orthopedic Knee-Ligament Injury, Anal. Chem. (2020). https://doi.org/10.1021/acs.analchem.0c03456.

[48] J. Liang, L. Bai, C. Dang, F. Cao, The K-type algorithms versus imbalanced data distributions, IEEE Trans. Fuzzy Syst. 20 (2012) 728–745. https://doi.org/10.1109/TFUZZ.2011.2182354.

[49] P. Fränti, S. Sieranoja, K-means properties on six clustering benchmark data sets, Appl. Intell. 48 (2018) 4743–4759. https://doi.org/10.1007/s10489-018-1238-7.

[50] L.C. Marney, W. Christopher Siegler, B.A. Parsons, J.C. Hoggard, B.W. Wright, R.E. Synovec, Tile-based Fisher-ratio software for improved feature selection analysis of comprehensive two-dimensional gas chromatography-time-of-flight mass spectrometry data, Talanta 115 (2013) 887–895. https://doi.org/10.1016/j.talanta.2013.06.038.

[51] B.A. Parsons, L.C. Marney, W.C. Siegler, J.C. Hoggard, B.W. Wright, R.E. Synovec, Tile-Based Fisher Ratio Analysis of Comprehensive Two-Dimensional Gas Chromatography Time-of-Flight Mass Spectrometry (GC × GC-TOFMS) Data Using a Null Distribution Approach, Anal. Chem. 87 (2015) 3812–3819. https://doi.org/10.1021/ac504472s.

Chapter 6: Enhancing Partial Least Squares Modeling of Comprehensive Two-Dimensional Gas Chromatography Time-of-Flight Mass Spectrometry Data by Tile-Based Variance Ranking

6.1. Introduction

The chemical composition of kerosene-based aerospace fuels directly impacts the reliability, reusability, and operability of rocket and jet engines. For example, differences in the original feedstock composition and/or production pathways may result in fuels with varying formulations despite belonging to the same fuel type (e.g., RP-1, RP-2, JP-5, Jet-A) [1–4]. Additionally, the presence of heteroatom or unsaturated species can detrimentally impact fuel thermal stability and generate carbonaceous deposits inside engine systems [5–9]. Therefore, strict specifications on fuel composition, and their physicochemical behavior in turn, are designated to ensure optimal and consistent performance. Comprehensive two-dimensional (2D) gas chromatography (GC×GC) is a powerful separation tool for the characterization of aerospace fuel composition [10–23]. Relative to its one-dimensional GC counterpart, GC×GC provides increased resolving power [24], selectivity [25], and sensitivity [26]. The use of an information-rich multivariate detector, such as time-of-flight mass spectrometry (GC×GC-TOFMS), further enhances analyte resolution and identification. However, sifting through the large amount of data produced with GC×GC-TOFMS to uncover compositional differences in numerous fuels requires advanced computational methods, referred to as chemometrics.

For fuel analysis, partial least squares (PLS) regression is a popular chemometric method for correlating chemical composition to measured physicochemical behavior. Mathematically,

---

This chapter is reproduced from C. N. Cain, G. S. Ochoa, R. E. Synovec, Enhancing Partial Least Squares Modeling of Comprehensive Two-Dimensional Gas Chromatography Time-of-Flight Mass Spectrometry Data by Tile-Based Variance Ranking, J. Chromatogr. A 1694 (2023), 463920.

PLS generates a regression model using a subset of features within the chromatograms that best predict the differences observed in the measured physical and/or chemical property [27]. The subset of features discovered to account for the variance between the chromatographic information and measured properties are termed latent variables (LVs). Two key outputs from the PLS model are the regression plot and linear regression vector (LRV). The regression plot highlights the linear relationship between the measured property values and those predicted by the PLS model while the LRV describes how features in the chromatograms are related to the physical and/or chemical property of interest. PLS regression has been beneficial in developing property-composition models to predict bulk content of hydrocarbons [17,22,28], fuel properties (e.g., density, viscosity, heat of combustion) [12,13,17,20,23,28], distillation curves [16], and fouling tendencies [18].

Prior to PLS analysis, it is highly advantageous that the GC×GC-TOFMS data be reduced to improve chemometric performance since the inclusion of noise and/or irrelevant signals can increase prediction errors and decrease computational speed [29]. Peak table generation and binning using a single-grid scheme are common data reduction strategies for GC×GC chromatograms. Peak tables generated for PLS analysis typically contain the total integrated signal for different hydrocarbon classes at different carbon numbers [12,20,22,23]. However, these tables may not include analytes with a low signal-to-noise ratio (*S/N*) [30], such as heteroatom species, that can influence a given fuel property. The presence of these missing values in the peak table creates an unstable PLS model [31]. On the other hand, binning commonly involves averaging the signal along both separation axes using a single-grid scheme of adjacent bins originating at the start of the chromatogram [13,16,17,28]. This approach ensures that the entire separation is utilized during PLS analysis; yet the bin size must be

carefully selected such that the compositional differences between fuel samples are still discernable for chemometric modeling [32]. Even with the appropriate bin size, detecting specific analytes influencing the model can be hindered if the analyte signal is split among adjacent bins [33].

Feature selection presents an alternative approach to GC×GC data reduction prior to PLS analysis. These methods remove uninformative signals from the data set by discovering analytes that indicate meaningful chemical variation between the samples [29]. Feature selection can be performed in either a supervised or unsupervised approach, where supervision refers to methods that utilize sample class information from the experimental design. Furthermore, feature selection can be performed using either a pixel-based, peak table-based, or tile-based approach [34]. Previously, Abrahamsson et al. demonstrated an improvement in PLS modeling of fouling tendencies for gas condensate samples after implementation of a supervised feature selection algorithm, termed RReliefF analysis [18]. Using a pixel-based approach, the RReliefF algorithm [35–37] weighted each data point in the chromatogram based on its correlation to reactor fouling. The data points with the largest positive weights, indicating high correlation to reactor fouling, were then used to construct their PLS model [18]. However, pixel-based feature selection methods can also obscure the discovery of true compositional differences due to the detection of instrumental artifacts from spurious detector noise and retention time misalignment [33].

Tile-based feature selection prevails among discovery-based analyses due to its advantages over pixel- and peak table-based methods [34], but its utility in reducing GC×GC data prior to PLS analysis has yet to be explored. Briefly, the tile-based approach was originally developed to discover class-distinguishing analytes by Fisher ratio (F-ratio) analysis [33,38]. The tile-based approach divides (i.e., tiles) the chromatogram into four spatially offset but

overlapping grid schemes, which ensures that every peak is optimally captured (i.e., roughly centered) within one of tiles. For each tile, the encapsulated chromatographic signal is then summed together on a per-mass channel (*m/z*) basis and an F-ratio, defined as the ratio of the between-class variance to the within-class variance, is calculated. The tile-based platform then automatically removes the redundant hits produced from the four overlapping grid schemes to ensure that each analyte is represented by a single entry in the hit list. The hit list ranks the features discovered in descending order of their F-ratio, where analyte hits near the top are more likely to be indicative of compositional differences between classes. Ultimately, the tiling platform provides three distinct advantages compared to pixel- and peak table-based feature selection methods: (1) mitigation of retention time shifting without the need for an alignment algorithm, (2) enhancements in the *S/N*, and (3) automated removal of redundant hits [33,38]. As a result, tile-based F-ratio analysis has been widely utilized in various studies, ranging from metabolomics to fuel analysis, and released into commercial software packages [15,39–43].

However, use of tile-based F-ratio analysis can be limited for experimental designs that do not have multiple samples/replicates per-class or knowledge of sample class membership. Therefore, the tile-based software has recently been extended to discover compositional differences in both pairwise [44] and unsupervised [45] analyses. Specifically, tile-based variance ranking was introduced as an unsupervised feature selection method to discover meaningful chemical differences without the need for class labels [45]. This method calculates the relative signal variance, defined as the square of the relative standard deviation (RSD^2), for every tile on a per-*m/z* basis. The total signal variance (RSD^2_{T}) for a tile can be expressed as

$$RSD^2{}_{\mathrm{T}} = RSD^2{}_{\mathrm{Chem}} + RSD^2{}_{\mathrm{B}} \tag{6.1}$$

where RSD^2_{Chem} and RSD^2_B denote the chemically relevant and background variance, respectively [45,46]. The chemically relevant variation (RSD^2_{Chem}) describes the true compositional differences between samples while the background variation (RSD^2_B) explains the uncertainty introduced from sample preparation and instrumentation, which can be minimized through proper normalization [45,46]. Hence, features with a larger relative signal variance (RSD^2_T) can be inferred to have more chemically meaningful variation (RSD^2_{Chem}). Tile-based variance ranking was previously coupled to two unsupervised chemometric methods, principal components analysis (PCA) and *k*-means clustering, to discover and identify compositional differences between three jet fuels [45].

Herein, this report establishes the first implementation of tile-based variance ranking, using RSD^2_T as defined in Eq. 6.1, as a selective, data reduction strategy to improve PLS modeling of GC×GC-TOFMS data. The tile-based variance ranking algorithm provides an automated method for discovering and identifying sample-related differences, which can be used to directly inform property-composition models. The advantages of this methodology are demonstrated using an extensive GC×GC-TOFMS data set of compositionally diverse aerospace fuels. PLS models for viscosity, hydrogen content, and heat of combustion are developed with the features discovered by tile-based variance ranking. Also, these features are subsequently analyzed with RReliefF to further reduce the hit list down to only the analytes that correlate with each fuel properties. Models constructed after tile-based variance ranking are compared to those developed using the standard, single-grid binning scheme for data reduction. This work demonstrates that tile-based variance ranking is a selective data reduction method for PLS modeling, providing lower prediction errors and direct identification of analytes in the LRVs.

6.2. Methods and Materials

6.2.1. Fuel data set

Previously, a set of 74 aerospace fuels were acquired from the Air Force Research Laboratory (AFRL, Edwards AFB, CA and Wright-Patterson AFB, OH) and analyzed by GC×GC-TOFMS [13]. This data set was comprised of fuels with both natural diversity from differences in their feedstock/production source and chemical diversity from blending several chemical streams. However, a small fraction of the fuels in this data set were previously observed to have an "atypical" composition (i.e., intense, overloaded peaks) and were demonstrated to not be modellable by PLS [13]. Therefore, these fuels were excluded from this study. The resulting fuel set was chemically diverse (Figure D.1), with samples comprised of both middle distillate products and blended formulations while also meeting different aerospace specifications (e.g., RP-1, RP-2, Jet-A, JP-5, JP-7, JP-8) [13]. Table 6.1 lists the 58 fuels used herein along with their measured physical properties. Note that the original sample numbers [13] were kept in this study for consistency. Viscosity and heat of combustion measurements were provided by AFRL (Edwards AFB, CA and Wright-Patterson AFB, OH) while measurements for hydrogen content were supplied by the Air Force Petroleum Office Fuels Laboratory (Vandenberg AFB, CA). The following ASTM methods were used to measure the physical properties of each fuel: viscosity at 40 ºC – ASTM D445/446 [47], hydrogen content – ASTM D7171 [48], and heat of combustion – ASTM D4809 [49].

GC×GC-TOFMS separations of these fuels were collected using an Agilent 6890N GC (Agilent Technologies, Palo Alto, CA), a thermal modulator (LECO, St. Joseph, MI), and a Pegasus III TOFMS (LECO, St. Joseph, MI). A 1 μL aliquot of each fuel sample was injected into the inlet with a split ratio of 200:1. The inlet temperature was set at 275 ºC. A "reverse"

column GC×GC configuration was employed herein, which consisted of a polar Rxi-17Sil MS (29.5 m × 250 µm inner diameter × 0.25 µm film thickness; Restek, Bellefonte, PA) column in the first dimension (^{1}D) column, and a non-polar Rxi-1MS (1.5 m × 180 µm × 0.18 µm; Restek, Bellefonte, PA) column in the second dimension (^{2}D) column. Ultra-high purity helium (Grade 5, 99.999 %; Praxair, Seattle, WA) was operated at a constant flow rate of 2 mL/min. The ^{1}D oven temperature was initially held at 40 ºC for 1.5 min before increasing to 200 ºC at 5 ºC/min, where it was held for 1 min. The ^{2}D oven and modulator followed the same temperature program with a +12 ºC and +30 ºC offset, respectively. The modulation period was 3 s. The transfer line and TOFMS ion source were set at 285 ºC and 225 ºC, respectively. Since fuel composition was not known beforehand, a 10 s acquisition delay was selected to protect the TOFMS filaments in case a solvent was present in the samples. The TOFMS recorded mass spectra from *m/z* 35-334 at 100 Hz with an electron impact ionization voltage of 70 eV. Further information regarding the experimental details for this data set can be found in our previous publication [13].

Table 6.1. List of the 58 fuel samples and their respective physical properties which were used in this study. A total of 42 fuels were used to develop the PLS models and 16 fuels (marked with an asterisk) comprised the external validation set. The original sample numbers from Berrier et al. [13] were kept herein for consistency.

Sample Number	Sample Name	Type	Viscosity (cSt)	Hydrogen Content (mass %)	Heat of Combustion (Btu/lbm)
1	YA2921HW10	RP-2	1.609	14.202	18626
2	BG1121GP04	RP-1	1.696	14.184	18626
3	GRC/0-100 HEP	RP-1	1.612	14.187	18594
4	WC0721HW01	RP-2	1.760	14.368	18660
5 *	LB073009-05	RP-1	1.710	14.054	18555
6	ZI1521HW10	RP-1	1.675	14.130	18559
7	CG0721HW10	RP-2	1.645	14.215	18583
8 *	LB073009-08	RP-1	1.600	14.183	18580
9	BB0821HW10	RP-2	1.705	14.178	18556
10 *	LB080409-05	RP-1	1.663	14.160	18587
11	ZI2621HW01	RP-2	1.852	14.276	18603
12 *	ZJ1321GP01	RP-2	1.662	14.218	18608
13	RG3021LS06	UL-RP-1	1.593	14.242	18604
14	RG3021LS05	RP-TS-5	1.584	14.216	18602

15	POSF 3327	JP-7	1.537	14.536	18690
17 *	LB073009-02	RP-1	1.588	14.055	18568
18	B0112868	RP-1	1.472	14.125	18590
19 *	VI2621LS01	RP-1	1.593	14.210	18598
21 *	DC310925	RP-1	1.600	14.234	18619
22	DC310923	RP-1	1.623	14.211	18623
23	DB131013	RP-1	1.647	14.249	18644
26	CB1121HW10	RP-2	1.688	14.122	18546
27	EA130720	RP-1	1.652	13.966	18514
28	EB220705	RP-1	1.559	14.005	18538
29 *	CHC JP-5	JP-5	1.347	14.117	18594
30 *	LB080409-01	RP-1	1.654	14.179	18553
31 *	LB073009-01	RP-1	1.720	14.146	18559
32	A0072256	RP-1	1.433	14.057	18636
34 *	LB100413-40	RP-1	1.644	14.093	18582
35 *	LB073009-03	RP-1	1.652	14.157	18590
36 *	LB073009-10	RP-1	1.593	14.188	18563
37	SA1421LS03	UL-RP-1	1.603	14.213	18584
38	ED060739	RP-1	1.667	13.970	18497
39	CB1121HW10	RP-2	1.694	14.157	18526
40 *	LB073009-09	RP-1	1.608	14.160	18552
41 *	LB073009-06	RP-1	1.726	14.266	18592
42	XC2521HW10	RP-1	1.588	14.253	18582
43	ZK0821HW20	RP-2	1.661	14.184	18550
44	ZK2121HW10	RP-2	1.526	14.180	18550
45	CL11-2928	JP-8	1.344	13.779	18487
49	CA2021HW10	RP-2	1.645	14.188	18572
51 *	41910	RP-1	1.512	14.266	18602
52	B01001634-01	RP-1	1.620	14.263	18579
53	POSF 4751	JP-8	1.340	13.932	18394
54	POSF 10359	JP-8	1.330	13.988	18553
55	POSF 10314	JP-8	1.510	13.908	18462
56	POSF 10312	JP-8	1.370	13.751	18463
57	POSF 10316	JP-8	1.290	13.971	18522
59	POSF 10358	Jet-A	1.430	13.946	18459
60	POSF 10311	Jet-A	1.370	13.499	18450
61	POSF 10315	Jet-A	1.140	14.282	18623
62	POSF 10369	Jet-A	1.180	13.985	18533
63	POSF 10313	Jet-A	1.480	13.618	18488
64	POSF 10337	JP-5	1.370	13.852	18514
65	POSF 10325	Jet-A	1.310	13.920	18506
66	POSF 10264	JP-8	1.140	14.410	18584
67	POSF 10289	JP-5	1.570	13.580	18429
68	POSF 9698	JP-8	1.300	14.078	18530

6.2.2. Data analysis

All data analysis was performed using Matlab 2019b (Mathworks, Natick, MA). After the data was imported, the chromatograms were baseline corrected and normalized to the sum of the total ion current (TIC) signal. The fuel data set of 58 fuels was then divided into two groups: a calibration set of 42 fuels and an external validation set of 16 fuels, which are marked by an asterisk in Table 6.1. Using the 42 fuels in the calibration data set, tile-based variance ranking via Eq. 6.1 was performed with a tile size of 12 s by 300 ms (^{1}D × ^{2}D) and a cluster window size of 9 s × 200 ms. A *S/N* threshold of 10 was implemented to remove tiles with noisy signals and the hit list was ranked according to the top RSD^2 *m/z* [45,50]. Next, a "stitch" GC×GC chromatogram was constructed to visualize the analyte locations discovered by tile-based variance ranking [15]. This method is based on the principles of the enhanced TIC chromatogram [51]. For every hit, a 12 s × 300 ms tile centered on the peak at the top RSD^2 *m/z* is extracted from the fuel chromatogram that has the largest signal for the given hit. Next, the algorithm zeros out the noise below a *S/N* threshold of 10 within each extracted tile. These tiles are then auto-scaled to make each analyte hit the same height for improved visualization before the tiles are inserted into an empty "chromatogram" (i.e., an array of zeros with the same size as the chromatogram) at the original retention time location. Note that the signals are additive for overlapping tiles. This process can be performed for every hit in the hit list or solely for those of interest depending on the visualization needs.

PLS models were developed using the calibration fuel set after reduction via either single-grid binning as previously performed [13,16,17,28], tile-based variance ranking, or coupling tile-based variance ranking to RReliefF analysis. The original size of the data set at the raw, pixel-level was 42 samples × 300 ^{2}D data points × 600 ^{1}D data points × 300 *m/z*. For single-

grid binning, the chromatographic data was divided into 12 s × 300 ms (^{1}D × ^{2}D) sections with the grid scheme anchored at the origin. Single-grid binning resulted in a reduced data set size of 42 samples × 10 ^{2}D data points × 150 ^{1}D data points × 300 *m/z*. Prior to PLS modeling, the single-grid binned data was unfolded into a vectorized format. Meanwhile, the peak area for every hit discovered by tile-based variance ranking was quantified in all 42 fuel samples at the top RSD^2 *m/z*. These peak areas were either utilized directly in PLS analysis or were then subjected to analysis with RReliefF. For the latter case, RReliefF was implemented to determine the correlation between the quantified peak areas and a given physical property measurement [35–37]. The discovered features were then re-ranked according to the predictor "importance" weight, resulting in three different lists, one for each physical property. The predictor importance weight for an analyte is based on the probability that two samples with different peak areas will have different physical property values [35–37]. Subsequent PLS models were constructed using only the peak areas for the top features designated in these re-ranked hit lists.

The calibration data from these reduction strategies were submitted to PLS Toolbox 8.9 (Eigenvector Research, Manson, WA) to build predictive models of viscosity, hydrogen content, and heat of combustion. The chromatographic data was mean-centered while the physical property measurements were auto-scaled. Venetian blinds cross-validation was implemented to determine the number of LVs to keep in each model and evaluate its predictive capability. For this study, this internal cross-validation procedure builds a calibration model using 35 of the 42 total fuel samples. Then, this sub-model is used to predict the values for the 7 fuels that were excluded. This process was repeated 6 times to determine the root-mean-square error of cross-validation (RMSECV), which is calculated as

$$RMSECV = \left[\frac{1}{N}\sum\left(y_{\mathrm{i,CV}} - y_{\mathrm{i,meas}}\right)^2\right]^{0.5} \tag{6.2}$$

where N is the number of fuel samples (i.e., 42) while $y_{i,CV}$ is the cross-validation predicted value and $y_{i,meas}$ is the measured value of sample i in the calibration data set. The RMSECV results were then normalized (NRMSECV) by the range of the measured values [13,17,52],

$$NRMSECV = \frac{RMSECV}{y_{meas,max} - y_{meas,min}} \times 100 \quad (6.3)$$

LRVs were also examined to discover compounds and/or hydrocarbon classes that were positively and negatively correlated with a given physical property. For cases where the analyte was poorly resolved, parallel factor analysis (PARAFAC) [53] was employed to acquire a high quality mass spectrum for analyte identification. Next, the chromatographic data from the external validation set of 16 fuel samples was inputted into the PLS models for further testing of their predictive ability. The RMSE of prediction (RMSEP) and normalized RMSEP (NRMSEP) were calculated as follows [52]:

$$RMSEP = \left[\frac{1}{N}\sum\left(y_{i,pred} - y_{i,meas}\right)^2\right]^{0.5} \quad (6.4)$$

$$NRMSEP = \frac{RMSEP}{y_{meas,max} - y_{meas,min}} \times 100 \quad (6.5)$$

where N is the number of fuel samples in the external validation set (i.e., 16), $y_{i,pred}$ is the predicted value of sample i from PLS and $y_{i,meas}$ is the measured value of sample i.

6.3. Results and Discussion

Figure 6.1 shows the TIC chromatograms for four aerospace fuels evaluated herein. A reverse column configuration, consisting of a polar ^{1}D column and a non-polar ^{2}D column, was selected for its improved resolution of these complex fuels [25]. This column set nominally separates the analytes by boiling point along the ^{1}D time axis and by polarity along the ^{2}D time axis. Hence, three distinct bands representing the different types of hydrocarbons in the fuel sample can be observed: alkanes (top; 2-3 s on ^{2}D), cycloalkanes (middle; 1-2 s on ^{2}D), and

aromatics (bottom; 0-1 s on ^{2}D). The approximate location of these bands, in turn, allow for broad interpretations regarding the chemical composition of each fuel. For instance, Sample 1 (RP-2) and Sample 32 (RP-1) are predominately composed of midweight alkanes and cycloalkanes (Figure 6.1A-B). In contrast, Sample 59 (Jet-A) and Sample 67 (JP-5) have a higher concentration of aromatics and are composed of a larger range of carbon numbers (Figure 6.1C-D). As illustrated by Figure 6.1 and Table 6.1, the compositional diversity of these fuels ultimately contributes to their differences in their physicochemical properties. Thus, use of chemometrics can provide a deeper insight into the link between fuel composition and performance.

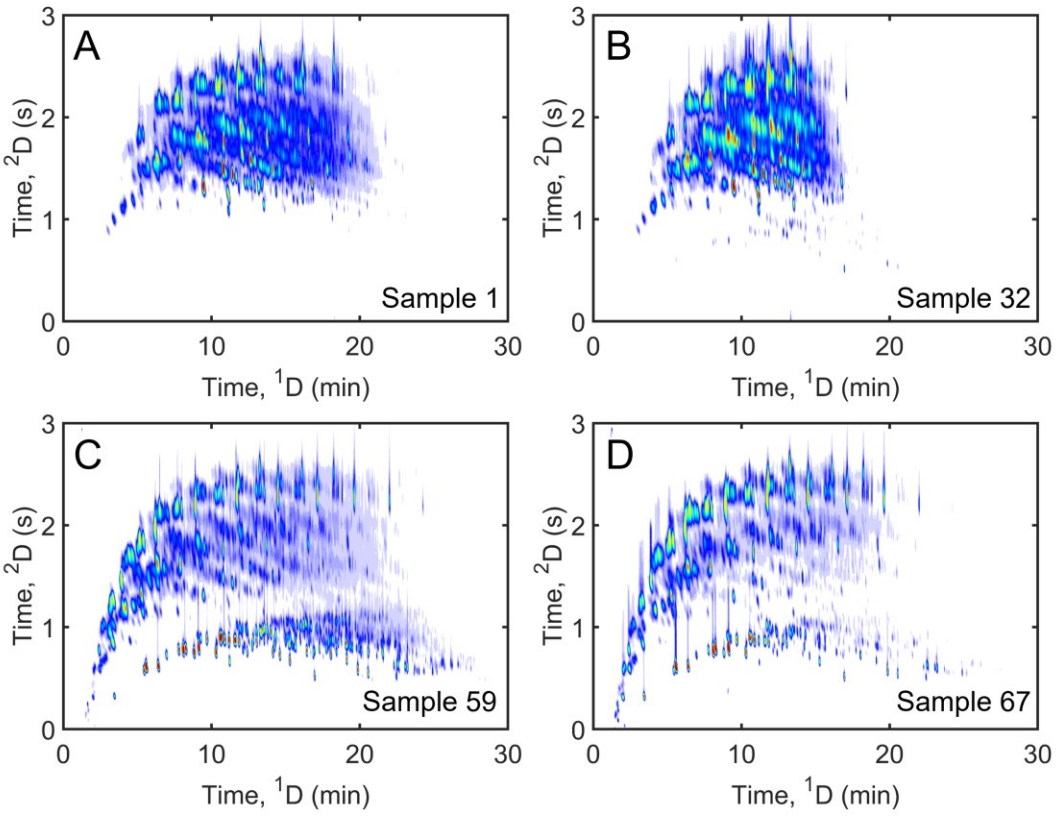


Figure 6.1. Total ion current (TIC) GC×GC chromatograms of four representative aerospace fuels analyzed in this study: (A) Sample 1 - RP-2, (B) Sample 32 - RP-1, (C) Sample 59 - Jet-A, and (D) Sample 67 - JP-5.

PLS regression was selected to correlate the compositional information hidden in the GC×GC-TOFMS data to viscosity, hydrogen content, and heat of combustion. Initially, PLS models for these three fuel properties were developed after binning the chromatographic data to a single-grid scheme, a common approach for data reduction. Figure 6.2A-C displays the respective regression plots for viscosity, hydrogen content, and heat of combustion for both the calibration (black unfilled circles) and external validation (blue filled diamonds) data sets. In general, a PLS model with a NRMSECV less than 10 % can be considered a good fit [13,52]. The NRMSECV for the PLS models using the single-grid binned data were slightly higher with values of 14.2 % for viscosity, 12.1 % for hydrogen content, and 14.4 % for heat of combustion. The NRMSEP for the validation subset were similar, with 14.3 % for viscosity, 11.0 % for hydrogen content, and 13.6 % for heat of combustion. It is also important to note that samples near the minimum or maximum limits measured for each property have the largest cross-validation residuals because of the chromatographic noise included in the model [54]. Also, Figure 6.2D-F shows the respective LRVs for viscosity, hydrogen content, and heat of combustion projected onto the dimensions of a GC×GC chromatogram. Chemically, the bins with positive values for the LRV (blue) correlate to an increase in a property measurement while bins with negative values (red) correlate to a decrease in that property. Note that binning ultimately reduces the chromatographic resolution and can sum away smaller compositional differences [32]. Furthermore, since peak signals can be split between bins, specific analytes that greatly influence the PLS model cannot be clearly identified. As a result, only broad generalizations about the influence of the boiling point range and different hydrocarbon compounds on each property can be made. The LRVs in Figure 6.2D-F highlight that compounds with a higher boiling point range were correlated with larger values for viscosity, hydrogen

content, and heat of combustion. Alkanes and cycloalkanes are correlated with higher viscosities, hydrogen content values, and a larger heat of combustion (Figure 6.2D-F). Conversely, aromatics are correlated with lower viscosities, hydrogen content values, and smaller heat of combustions (Figure 6.2D-F). Hence, the influence of individual chemical species, especially those with low signal, on these properties cannot be addressed in these models.

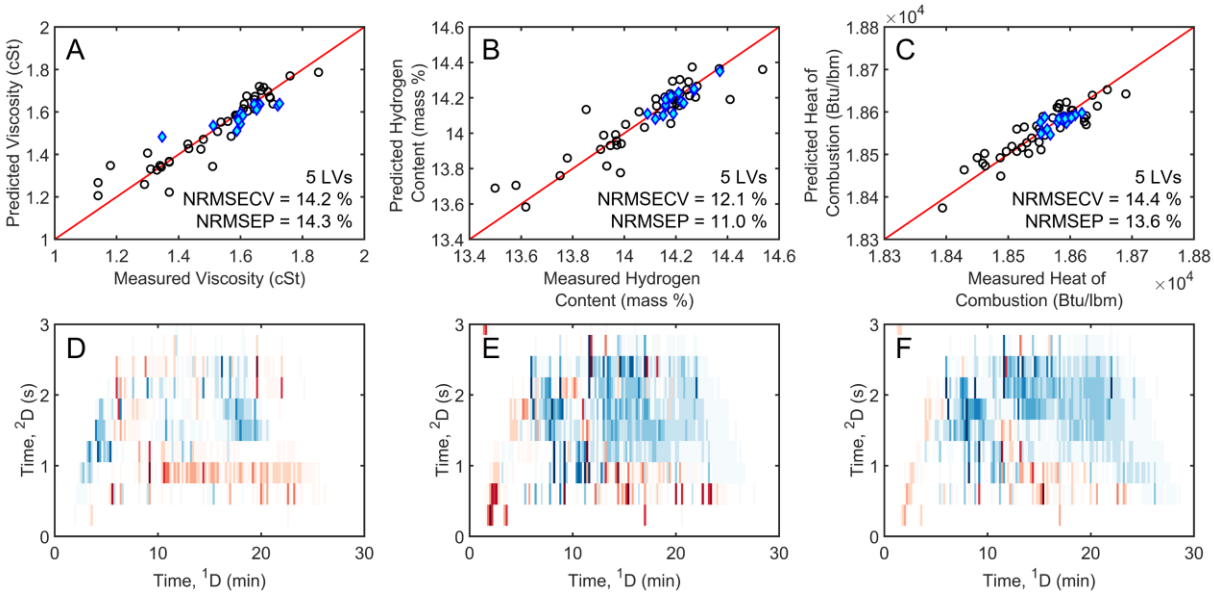


Figure 6.2. PLS prediction of viscosity (left), hydrogen content (middle), and heat of combustion (right) using single-grid binning of GC×GC-TOFMS data. (A-C) Regression plots for each physical property. The red line represents ideal agreement between the predicted and measured values. Samples used to build the calibration model are shown as black unfilled circles while samples used in the external validation set are shown as blue filled diamonds. The number of LVs, NRMSECV, and NRMSEP for each PLS model is provided. (D-F) LRVs generated for each physical property, where positively loaded values are highlighted in blue and negatively loaded values are shown in red. All LRVs are plotted on the same color scale.

To develop a more accurate and sensitive PLS model for these three fuel properties, tile-based variance ranking was implemented to discover compositional differences between the chromatograms. The advantages of this approach are that retention time misalignment is mitigated prior to feature selection and chromatographic differences can be discovered without the need for multiple injection replicates per-sample [33,38,45]. In total, 521 hits were

discovered by tile-based variance ranking (Figure 6.3). Figure 6.3A plots the logarithm of the RSD^2 versus the logarithm of the summed ^{2}D peak area measured at the top RSD^2 *m/z* (red) and all secondary *m/z* (blue) for every hit. This plot confirms that all analytes discovered have ample signal at the top RSD^2 *m/z*. The distribution of RSD^2 measured at the top *m/z* for the 521 hits is shown in Figure 6.3B. The RSD^2 ranges from 0.07 to 22.84 (i.e., an *RSD* between 26 % and 478 %), which demonstrates that tile-based variance ranking can discover analytes that were present in only a few samples and exhibited minor changes in intensity between samples. The stitch GC×GC chromatogram in Figure 6.3C visualizes the locations for the 521 hits discovered by tile-based variance ranking, whereby the 2D signal at the top RSD^2 *m/z* for each hit has been stitched onto an empty 2D background. This stitch chromatogram illustrates that both low boiling point alkanes and aromatics are mainly responsible for the compositional differences between the fuels. These compositional differences can also be observed in the four fuels shown in Figure 6.1, demonstrating the utility of the stitch chromatogram visualization tool.

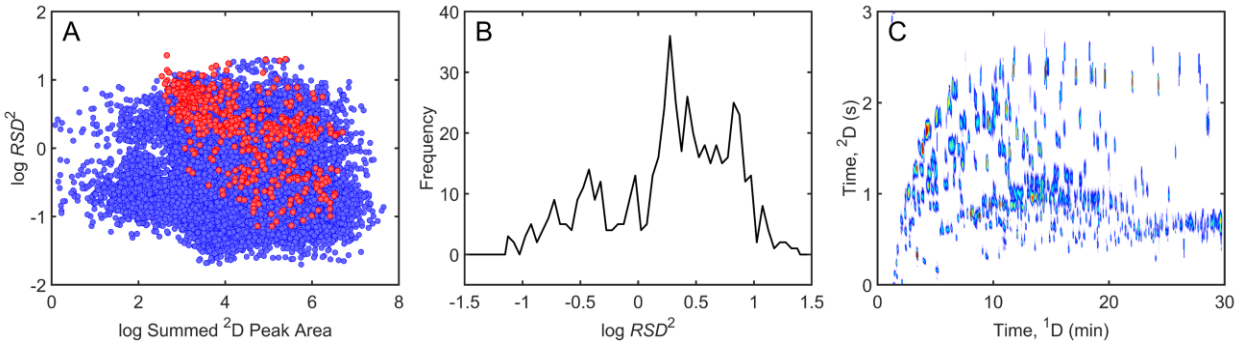


Figure 6.3. Summary of tile-based variance ranking results. (A) Plot of the log RSD^2 versus log of the summed ^{2}D peak area measured per-*m/z* for all 521 hits. Red dots emphasize the results for the top RSD^2 *m/z* and blue dots are the results for all other *m/z* detected for each hit. (B) Distribution of the log RSD^2 calculated at the top *m/z* for each hit discovered. (C) Stitch GC×GC chromatogram of the 521 hits discovered. The stitch chromatogram was constructed by pulling the data at the top RSD^2 *m/z* from the fuel with the largest signal for a given hit.

Using the summed ^{2}D peak area at the top RSD^2 *m/z* for all 521 hits (summed signal for each individual analyte peak in the stitch chromatograms), PLS models were developed for the

three fuel properties of interest (Figure 6.4). Figure 6.4A-C illustrates the regression plots for viscosity (NRMSECV = 10.5 %), hydrogen content (NRMSECV = 8.3 %), and heat of combustion (NRMSECV = 13.1 %). The NRMSEP for viscosity, hydrogen content, and heat of combustion using the external fuel set (blue diamonds) was 10.2 %, 7.6 %, and 13.5 %, respectively. Compared to the initial regression plots shown in Figure 6.2A-C, the accuracy of the models improved, which is a consequence of selectively utilizing the information hidden in the chromatographic data. For instance, the single-grid binning scheme implemented in Figure 6.2A-C reduces each chromatogram along the separation axes (e.g., 180,000 un-binned data points down to 1,500 bins), but not along the *m/z* dimension. Hence, noisy *m/z* along with irrelevant signals were still included in the PLS models shown in Figure 6.2A-C. Conversely, tile-based variance ranking selectively utilizes both the chromatographic and *m/z* dimensions to provide feature selection with data reduction. Specifically, the peak areas for the 521 features with relevant sample-related differences that were discovered by tile-based variance ranking became the 521 data points for each fuel sample. As a result, the accuracy of the PLS models improved after application of tile-based variance ranking, which is demonstrated by the smaller residuals between the predicted and measured property values. Furthermore, Figure 6.4D-F shows the LRVs for each property. For visualization of the LRVs as a GC×GC chromatograms, the values from the PLS model were projected onto the tile dimensions surrounding each discovered analyte. Generally, the chemical interpretation for the LRVs shown in Figure 6.4D-F is similar to the results observed previously (Figure 6.2D-F). Figure 6.4D-F shows that the alkanes and cycloalkane regions are positively correlated with viscosity, hydrogen content, and heat of combustion while the aromatics region was negatively correlated. Notably, the use of tile-based variance ranking also allows the impact of specific analytes on the PLS model to be

determined (Table D.1 - Table D.3). For example, Figure 6.4D interestingly shows that the presence of 1-eicosanol was correlated to samples with higher viscosities. This finding is in agreement with previous work, which showed that saturated fatty acids increase fuel viscosity and can lead to the formation of carbon deposits inside of engines [55]. Likewise, long chain alkanes like tetradecane were correlated with larger measurements for hydrogen content and heat of combustion while substituted naphthalenes and cycloalkanes were correlated with lower values for hydrogen content and heat of combustion (Figure 6.4E-F). Ultimately, Figure 6.4 demonstrates that tile-based variance ranking improved model quality relative to using a single-grid bin scheme for the three physical properties due to the large reduction of noisy/irrelevant features from the data.

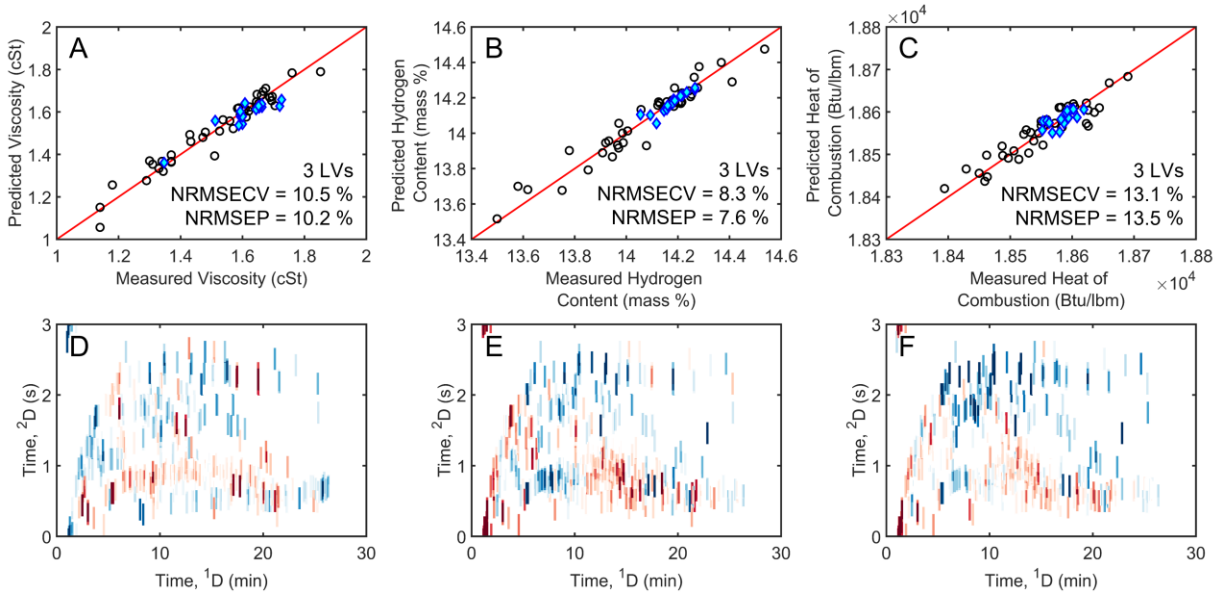


Figure 6.4. PLS prediction of viscosity (left), hydrogen content (middle), and heat of combustion (right) using the features discovered by tile-based variance ranking. (A-C) Regression plots for each physical property. The red line represents ideal agreement between the predicted and measured values. Samples used to build the calibration model are shown as black unfilled circles while samples used in the external validation set are shown as blue filled diamonds. The number of LVs, NRMSECV, and NRMSEP for each PLS model is provided. (D-F) LRVs generated for each physical property, where positively loaded values are highlighted in blue and negatively loaded values are shown in red. All LRVs are plotted on the same color scale.

While tile-based variance ranking provides a means for feature selection with data reduction, the hit list can be further distilled down to contain only those analytes that are highly correlated with each property. In turn, the predictive capabilities of the PLS model will improve, which will be shown with smaller values for the NRMESCV and NRMSEP. A simple solution would be the use of an RSD^2 threshold, but since the magnitude of the RSD^2 does not imply an association to a given fuel property, model performance can deteriorate if the threshold is set too high or low (Figure D.2; Figure D.3). Herein, the RReliefF algorithm was explored to discover features in the hit list that were correlated with each physical property (Figure 6.5). In the context of this study, this machine learning algorithm weights the discovered features based on their ability to distinguish between samples [35–37]. A positive weight will be given to an analyte in the hit list if its signal and the property measurements both have large variations between the samples [18]. Conversely, a negative weight will be assigned if the analyte signal has a large variation between samples, but the measured property does not change [18]. For each of the three fuel properties modeled in this study, the RReliefF algorithm was used to weight the features discovered by tile-based variance ranking based on their correlation to the property measurements. Following this computation, the features in the hit list were re-ranked according to their importance weight, which describes how well each feature can discriminate between samples with different physical properties (Figure 6.5A-C). Notably, the RReliefF algorithm assigned negative values (i.e., importance weight < 0) to 60 features for viscosity, 64 features for hydrogen content, and 38 features for heat of combustion. Features with these negative weights were considered by the RReliefF algorithm to be highly irrelevant in predicting the physical properties. To further facilitate downstream modeling, a relevance threshold was determined by evaluating how the features selected affected model performance [56]. Herein, this threshold was

determined by finding the importance weight where the NRMSECV for the given PLS model is minimized (Figure 6.5D-F). The arrow with the yellow star in Figure 6.5 indicates the weight threshold determined for the three fuel properties. Thresholds for viscosity (importance weight > 0.053), hydrogen content (importance weight > 0.071), and heat of combustion (importance weight > 0.019) selected only 48, 125, and 172 of the original 521 features discovered by tile-based variance ranking.

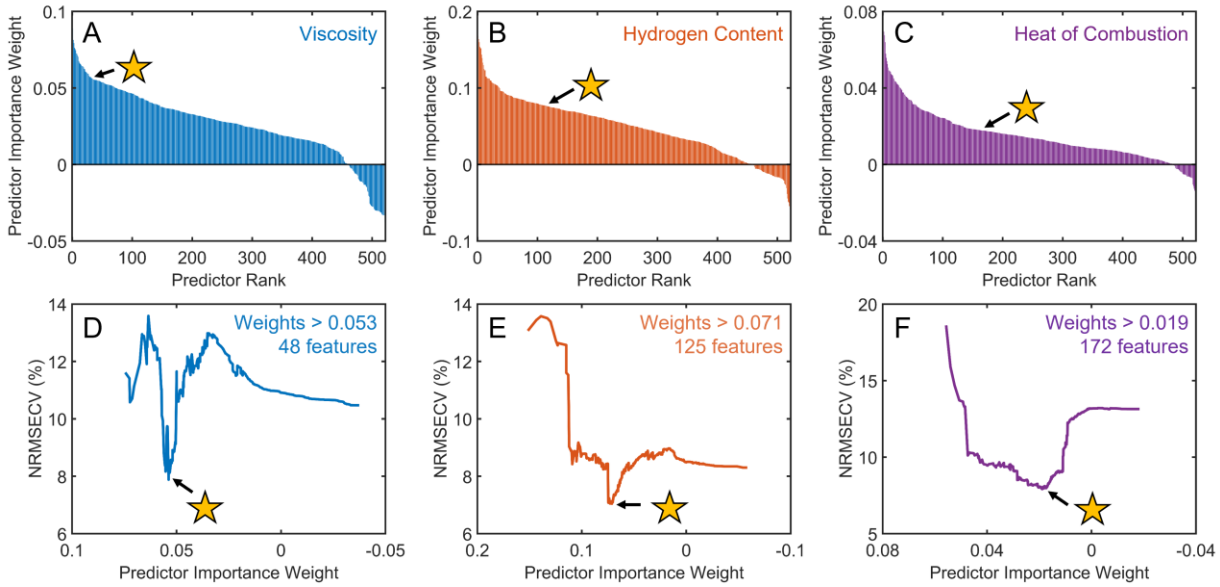


Figure 6.5. Ranking features discovered using the RReliefF algorithm in order of predictor importance from left to right for modeling viscosity (left), hydrogen content (middle), and heat of combustion (right). The arrow and yellow star indicate the number of features selected by RReliefF feature optimization to model each physical property. (A-C) Bar plots of predictor importance weight, which was used to re-rank the chromatographic features. (D-F) Selection of the predictor importance weight threshold based on the NRMSECV for each model.

Figure 6.6 shows the stitch chromatograms for the features that had an importance weight above the threshold determined for viscosity (A), hydrogen content (B), and heat of combustion (C). The stitch chromatogram for viscosity (Figure 6.6A) highlights the 48 features with an importance weight above 0.053, which equated to minimizing the PLS modeling error (Figure 6.5D). While the stitch chromatogram of the features selected for modeling viscosity is sparse, compounds from various types of hydrocarbons (e.g., alkanes, cycloalkanes, and aromatics) and

molecular weights were chosen (Figure 6.6A). In contrast, Figure 6.6B shows that a majority of the 125 compounds chosen to model hydrogen content (Figure 6.5E) were aromatics and heteroatom species. This result is attributed to the aromatics section showing the greatest diversity between the fuels in this data, which ultimately affects the measured hydrogen content [13]. Lastly, Figure 6.6C demonstrates that the 172 features selected by the RReliefF feature optimization method to model heat of combustion (Figure 6.5F) were mainly alkanes and aromatics. This selection is expected since previous work demonstrated that the combustion properties of aerospace fuels are strongly correlated with their aromaticity [14,57].

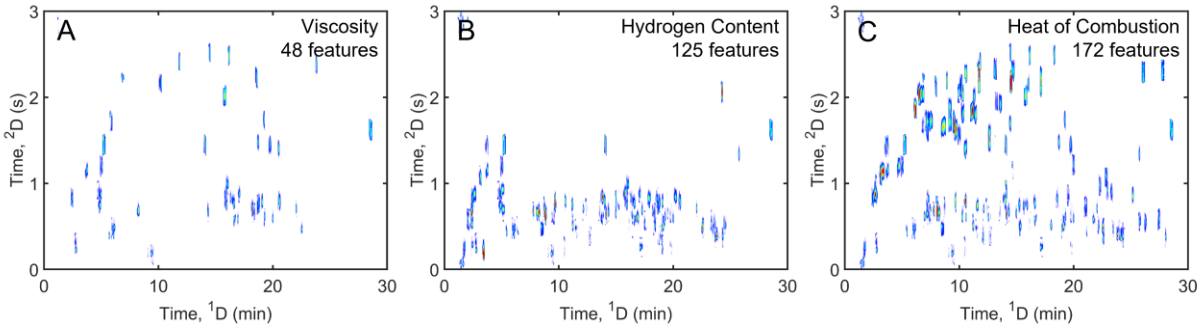


Figure 6.6. Stitch GC×GC chromatograms of the most important features for modeling viscosity (A), hydrogen content (B), and heat of combustion (C). The stitch chromatograms were constructed by pulling the data at the top RSD^2 *m/z* from the fuel with the largest signal for a given hit.

The final PLS models for viscosity, hydrogen content, and heat of combustion are shown in Figure 6.7. These models were built using the summed ^{2}D peak area at the top RSD^2 *m/z* for the subset of features selected by the RReliefF feature optimization algorithm to model each property (Figure 6.5; Figure 6.6). Figure 6.7A-C shows that the models built for the three physical properties all have a NRMSECV and NRMSEP less than 8.5 %. The similarity between the NRMSECV and NRMSEP for each physical property demonstrates the stability of the PLS models and their ability to fit fuels with a wide variety of chemical compositions, including blended fuels. For reference, the NRMSECV (NRMSEP) for models developed after single-grid

binning and initial feature selection with tile-based variance ranking ranged between 12.1-14.4 % (11.0-14.3 %) and 8.3-13.1 % (7.6-13.5 %), respectively (Figure 6.2; Figure 6.4). Thus, model performance was significantly improved by using tile-based variance ranking for initial feature selection with data reduction and RReliefF analysis as an additional feature optimization step. The LRVs shown in Figure 6.7D-F broadly show the same chemical interpretations as the previous LRVs (Figure 6.2; Figure 6.4); however, the effect of each analyte on the different physical properties can be readily identified (Table D.4 - Table D.6). For example, analytes with larger molecular weights like pristane, 1-eicosanol, hexylcyclohexane, and dihydro-(-)-neoclovene-(I) were positively correlated with viscosity (Figure 6.7D). Whereas lighter analytes, such as 1-methylcyclohexene and (*Z*)-undecene, were negatively correlated with viscosity (Figure 6.7D). These findings are consistent with previous studies showing that viscosity increases as the molecular size and weight increases [13,58]. The LRV for hydrogen content revealed that alkanes and alkenes like 1,3-dicyclohexylbutane, 1-methylcyclohexene, and bicyclo[2.2.2]octane were positively correlated (Figure 6.7E). Meanwhile, aromatics like 1,2,3-trimethylbenzene, toluene, and naphthalene were negatively correlated with hydrogen content (Figure 6.7E), which is expected given their lower hydrogen to carbon ratio. Interestingly, Figure 6.7F also reveals that many linear alkanes like dodecane, tridecane, and tetradecane are positively correlated with heat of combustion. Conversely, branched alkanes (e.g., 2,2,3,3-tetramethylbutane, 2,2-dimethylpentane, and 2-methylhexane) were negatively correlated with heat of combustion (Figure 6.7F). These modeling results support previous work demonstrating that the combustion behavior of linear and branched alkanes is different, where the former can produce more CO_2 through decomposition [59]. Collectively, Figure 6.7 demonstrates the

advantages of using tile-based analyses coupled with RReliefF feature selection optimization to improve and support property-composition studies.

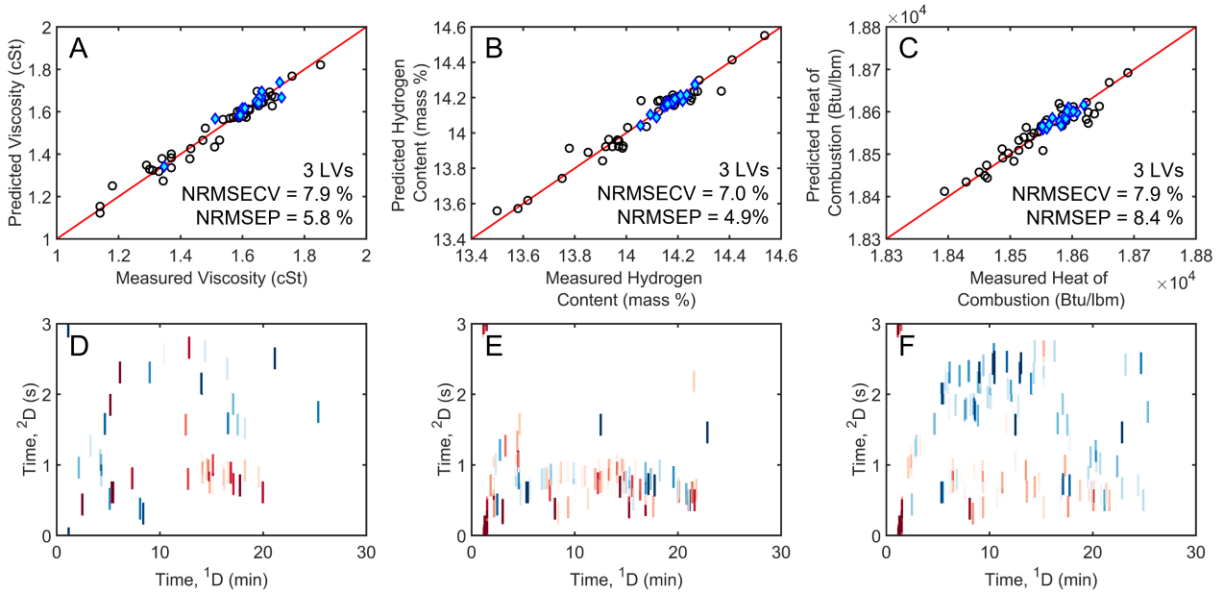


Figure 6.7. PLS prediction of viscosity (left), hydrogen content (middle), and heat of combustion (right) using the features with the highest importance as indicated by the RReliefF algorithm. (A-C) Regression plots for each physical property. The red line represents ideal agreement between the predicted and measured values. Samples used to build the calibration model are shown as black unfilled circles while samples used in the external validation set are shown as blue filled diamonds. The number of LVs, NRMSECV, and NRMSEP for each PLS model is provided. (D-F) LRVs generated for each physical property, where positively loaded values are highlighted in blue and negatively loaded values are shown in red. All LRVs are plotted on the same color scale.

6.4. Conclusion

This work illustrates a novel tile-based feature selection workflow for discovering differences in fuel chemical composition and connecting those differences to their performance properties. The NRMSECV (NRMSEP) for the initial PLS models constructed with the chromatographic data binned to a single-grid scheme was 14.2 % (14.3 %) for viscosity, 12.1 % (11.0 %) for hydrogen content, and 14.4 % (13.6 %) for heat of combustion. To improve the accuracy of these property-composition models, tile-based variance ranking was applied to provide both feature selection and data reduction. This unsupervised feature selection method

discovered 521 analytes that defined both large and minute differences in analyte concentration among the fuel samples ($0.07 < RSD^2 < 22.84$). Lower prediction errors (NRMSECV and NRMSEP) were observed for viscosity (10.5 % and 10.2 %), hydrogen content (8.3 % and 7.6 %), and heat of combustion (13.1 % and 13.5 %) using all 521 features discovered by tile-based variance ranking. The RReliefF algorithm was then applied to further refine and hence reduce the list of features to those that are strongly correlated to viscosity (48 analytes), hydrogen content (125 analytes), or heat of combustion (172 analytes). After applying the RReliefF feature optimization algorithm, the respective NRMSECVs for the final PLS models for viscosity, hydrogen content, and heat of combustion were 7.9 %, 7.0 %, and 7.9 %. Furthermore, the NRMSEPs for these final models were 5.8 % for viscosity, 4.9 % for hydrogen content, and 8.4 % for heat of combustion, demonstrating the stability of the models after RReliefF feature optimization with tile-based variance ranking. It is also important to note that this tile-based workflow allows for the identification of specific analytes that highly influence the model for a given fuel property. These results can be used to improve both routine fit-for-purpose testing and the design of new fuel formulations. Broadly, the presented methodology can be extended to different chromatographic applications and chemometric models like partial least squares-discriminant analysis (PLS-DA) to construct highly accurate and sensitive multivariate models.

6.5. References

[1] L.S. Ott, A.B. Hadler, T.J. Bruno, Variability of the rocket propellants RP-1, RP-2, and TS-5: Application of a composition- and enthalpy-explicit distillation curve method, Ind. Eng. Chem. Res. 47 (2008) 9225–9233. https://doi.org/10.1021/ie800988u.

[2] T.J. Bruno, M.L. Huber, E.W. Lemmon, Effect of RP-1 compositional variability on thermophysical properties, Energy Fuels 23 (2009) 5550–5555. https://doi.org/10.1021/ef900597q.

[3] T.M. Lovestead, B.C. Windom, J.R. Riggs, C. Nickell, T.J. Bruno, Assessment of the compositional variability of RP-1 and RP-2 with the advanced distillation curve approach, Energy Fuels 24 (2010) 5611–5623. https://doi.org/10.1021/ef100994w.

[4] T.M. Lovestead, J.L. Burger, N. Schneider, T.J. Bruno, Comprehensive Assessment of Composition and Thermochemical Variability by High Resolution GC/QToF-MS and the Advanced Distillation-Curve Method as a Basis of Comparison for Reference Fuel Development, Energy Fuels 30 (2016) 10029–10044. https://doi.org/10.1021/acs.energyfuels.6b01837.

[5] N.J. Kuprowicz, S. Zabarnick, Z.J. West, J.S. Ervin, Use of measured species class concentrations with chemical kinetic modeling for the prediction of autoxidation and deposition of jet fuels, Energy Fuels 21 (2007) 530–544. https://doi.org/10.1021/ef060391o.

[6] M. Sobkowiak, J.M. Griffith, B. Wang, B. Beaver, Insight into the mechanisms of middle distillate fuel oxidative degradation. Part 1: On the role of phenol, indole, and carbazole derivatives in the thermal oxidative stability of fischer-tropsch/petroleum jet fuel blends, Energy Fuels 23 (2009) 2041–2046. https://doi.org/10.1021/ef8006992.

[7] B. Jin, K. Jing, J. Liu, X. Zhang, G. Liu, Pyrolysis and coking of endothermic hydrocarbon fuel in regenerative cooling channel under different pressures, J. Anal. Appl. Pyrolysis 125 (2017) 117–126. https://doi.org/10.1016/j.jaap.2017.04.010.

[8] E. Alborzi, P. Gadsby, M.S. Ismail, A. Sheikhansari, M.R. Dwyer, A.J.H.M. Meijer, S.G. Blakey, M. Pourkashanian, Comparative Study of the Effect of Fuel Deoxygenation and Polar Species Removal on Jet Fuel Surface Deposition, Energy Fuels 33 (2019) 1825–1836. https://doi.org/10.1021/acs.energyfuels.8b03468.

[9] S. Zabarnick, Z.J. West, L.M. Shafer, S.S. Mueller, R.C. Striebich, P.J. Wrzesinski, Studies of the Role of Heteroatomic Species in Jet Fuel Thermal Stability: Model Fuel Mixtures and Real Fuels, Energy Fuels 33 (2019) 8557–8565. https://doi.org/10.1021/acs.energyfuels.9b02345.

[10] M.K. Jennerwein, M. Eschner, T. Gröger, T. Wilharm, R. Zimmermann, Complete group-type quantification of petroleum middle distillates based on comprehensive two-dimensional gas chromatography time-of-flight mass spectrometry (GCxGC-TOFMS) and visual basic scripting, Energy Fuels 28 (2014) 5670–5681. https://doi.org/10.1021/ef501247h.

[11] B. Kehimkar, J.C. Hoggard, L.C. Marney, M.C. Billingsley, C.G. Fraga, T.J. Bruno, R.E. Synovec, Correlation of rocket propulsion fuel properties with chemical composition using comprehensive two-dimensional gas chromatography with time-of-flight mass spectrometry followed by partial least squares regression analysis, J. Chromatogr. A 1327 (2014) 132–140. https://doi.org/10.1016/j.chroma.2013.12.060.

[12] P. Vozka, B.A. Modereger, A.C. Park, W.T.J. Zhang, R.W. Trice, H.I. Kenttämaa, G. Kilaz, Jet fuel density via GC × GC-FID, Fuel 235 (2019) 1052–1060. https://doi.org/10.1016/j.fuel.2018.08.110.

[13] K.L. Berrier, C.E. Freye, M.C. Billingsley, R.E. Synovec, Predictive Modeling of Aerospace Fuel Properties Using Comprehensive Two-Dimensional Gas Chromatography with Time-Of-Flight Mass Spectrometry and Partial Least Squares Analysis, Energy Fuels 34 (2020) 4084–4094. https://doi.org/10.1021/acs.energyfuels.9b04108.

[14] J. Heyne, D. Bell, J. Feldhausen, Z. Yang, R. Boehm, Towards fuel composition and properties from Two-dimensional gas chromatography with flame ionization and vacuum ultraviolet spectroscopy, Fuel 312 (2022) 122709. https://doi.org/10.1016/j.fuel.2021.122709.

[15] G.S. Ochoa, M.C. Billingsley, R.E. Synovec, Using solid-phase extraction to facilitate a focused tile-based Fisher ratio analysis of comprehensive two-dimensional gas chromatography time-of-flight mass spectrometry data: comparative analysis of aerospace fuel composition, Anal. Bioanal. Chem. (2022). https://doi.org/10.1007/s00216-022-04348-1.

[16] B. Kehimkar, B.A. Parsons, J.C. Hoggard, M.C. Billingsley, T.J. Bruno, R.E. Synovec, Modeling RP-1 fuel advanced distillation data using comprehensive two-dimensional gas chromatography coupled with time-of-flight mass spectrometry and partial least squares analysis, Anal. Bioanal. Chem. 407 (2015) 321–330. https://doi.org/10.1007/s00216-014-8233-6.

[17] C.E. Freye, B.D. Fitz, M.C. Billingsley, R.E. Synovec, Partial least squares analysis of rocket propulsion fuel data using diaphragm valve-based comprehensive two-dimensional gas chromatography coupled with flame ionization detection, Talanta 153 (2016) 203–210. https://doi.org/10.1016/j.talanta.2016.03.016.

[18] V. Abrahamsson, N. Ristic, K. Franz, K. Van Geem, Comprehensive two-dimensional gas chromatography in combination with pixel-based analysis for fouling tendency prediction, J. Chromatogr. A 1501 (2017) 89–98. https://doi.org/10.1016/j.chroma.2017.04.021.

[19] M.K. Jennerwein, A.C. Sutherland, M. Eschner, T. Gröger, T. Wilharm, R. Zimmermann, Quantitative analysis of modern fuels derived from middle distillates – The impact of diverse compositions on standard methods evaluated by an offline hyphenation of HPLC-refractive index detection with GC×GC-TOFMS, Fuel 187 (2017) 16–25. https://doi.org/10.1016/j.fuel.2016.09.033.

[20] X. Shi, H. Li, Z. Song, X. Zhang, G. Liu, Quantitative composition-property relationship of aviation hydrocarbon fuel based on comprehensive two-dimensional gas chromatography with mass spectrometry and flame ionization detector, Fuel 200 (2017) 395–406. https://doi.org/10.1016/j.fuel.2017.03.073.

[21] R.L. Webster, P.M. Rawson, C. Kulsing, D.J. Evans, P.J. Marriott, Investigation of the Thermal Oxidation of Conventional and Alternate Aviation Fuels with Comprehensive Two-Dimensional Gas Chromatography Accurate Mass Quadrupole Time-of-Flight Mass Spectrometry, Energy Fuels 31 (2017) 4886–4894. https://doi.org/10.1021/acs.energyfuels.7b00178.

[22] R. Chakravarthy, C. Acharya, A. Savalia, G.N. Naik, A.K. Das, C. Saravanan, A. Verma, K.B. Gudasi, Property Prediction of Diesel Fuel Based on the Composition Analysis Data by two-Dimensional Gas Chromatography, Energy Fuels 32 (2018) 3760–3774. https://doi.org/10.1021/acs.energyfuels.7b03822.

[23] P. Vozka, H. Mo, P. Šimáček, G. Kilaz, Middle distillates hydrogen content via GC×GC-FID, Talanta 186 (2018) 140–146. https://doi.org/10.1016/j.talanta.2018.04.059.

[24] M.S. Klee, J. Cochran, M. Merrick, L.M. Blumberg, Evaluation of conditions of comprehensive two-dimensional gas chromatography that yield a near-theoretical maximum in peak capacity gain, J. Chromatogr. A 1383 (2015) 151–159. https://doi.org/10.1016/j.chroma.2015.01.031.

[25] M. Jennerwein, M. Eschner, T. Wilharm, T. Gröger, R. Zimmermann, Evaluation of reversed phase versus normal phase column combination for the quantitative analysis of common commercial available middle distillates using GC × GC-TOFMS and Visual Basic Script, Fuel 235 (2019) 336–338. https://doi.org/10.1016/j.fuel.2018.07.081.

[26] A.L. Lee, K.D. Bartle, A.C. Lewis, A model of peak amplitude enhancement in orthogonal two-dimensional gas chromatography, Anal. Chem. 73 (2001) 1330–1335. https://doi.org/10.1021/ac001120s.

[27] S. Wold, M. Sjöström, L. Eriksson, PLS-regression: A basic tool of chemometrics, Chemom. Intell. Lab. Syst. 58 (2001) 109–130. https://doi.org/10.1016/S0169-7439(01)00155-1.

[28] B. Kehimkar, J.C. Hoggard, L.C. Marney, M.C. Billingsley, C.G. Fraga, T.J. Bruno, R.E. Synovec, Correlation of rocket propulsion fuel properties with chemical composition using comprehensive two-dimensional gas chromatography with time-of-flight mass spectrometry followed by partial least squares regression analysis, J. Chromatogr. A 1327 (2014) 132–140. https://doi.org/10.1016/j.chroma.2013.12.060.

[29] T. Mehmood, K.H. Liland, L. Snipen, S. Sæbø, A review of variable selection methods in Partial Least Squares Regression, Chemom. Intell. Lab. Syst. 118 (2012) 62–69. https://doi.org/10.1016/j.chemolab.2012.07.010.

[30] H.D. Bean, J.E. Hill, J.M.D. Dimandja, Improving the quality of biomarker candidates in untargeted metabolomics via peak table-based alignment of comprehensive two-dimensional gas chromatography-mass spectrometry data, J. Chromatogr. A 1394 (2015) 111–117. https://doi.org/10.1016/j.chroma.2015.03.001.

[31] B. Walczak, D.L. Massart, Dealing with missing data: Part I, Chemom. Intell. Lab. Syst. 58 (2001) 15–27. https://doi.org/10.1016/S0169-7439(01)00131-9.

[32] P.E. Sudol, D. V. Gough, S.E. Prebihalo, R.E. Synovec, Impact of data bin size on the classification of diesel fuels using comprehensive two-dimensional gas chromatography with principal component analysis, Talanta 206 (2020) 120239. https://doi.org/10.1016/j.talanta.2019.120239.

[33] L.C. Marney, W. Christopher Siegler, B.A. Parsons, J.C. Hoggard, B.W. Wright, R.E. Synovec, Tile-based Fisher-ratio software for improved feature selection analysis of comprehensive two-dimensional gas chromatography-time-of-flight mass spectrometry data, Talanta 115 (2013) 887–895. https://doi.org/10.1016/j.talanta.2013.06.038.

[34] T.J. Trinklein, C.N. Cain, G.S. Ochoa, S. Schöneich, L. Mikaliunaite, R.E. Synovec, Recent Advances in GC×GC and Chemometrics to Address Emerging Challenges in Nontargeted Analysis, Anal. Chem. 95 (2023) 264–286. https://doi.org/10.1021/acs.analchem.2c04235.

[35] M. Robnik-Šikonja, I. Kononenko, An adaptation of Relief for attribute estimation in regression, Mach. Learn. Proc. Fourteenth Int. Conf. 5 (1997) 296–304.

[36] M. Robnik-Šikonja, I. Kononenko, Theoretical and Empirical Analysis of ReliefF and RReliefF, Mach. Learn. 53 (2003) 23–69. https://doi.org/https://doi.org/10.1023/A:1025667309714.

[37] K.C. Patchava, M. Benaissa, H. Behairy, Improving the prediction performance of PLSR using RReliefF and FSD for the quantitative analysis of glucose in Near Infrared spectra, Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. EMBS (2015) 2379–2382. https://doi.org/10.1109/EMBC.2015.7318872.

[38] B.A. Parsons, L.C. Marney, W.C. Siegler, J.C. Hoggard, B.W. Wright, R.E. Synovec, Tile-Based Fisher Ratio Analysis of Comprehensive Two-Dimensional Gas Chromatography Time-of-Flight Mass Spectrometry (GC × GC–TOFMS) Data Using a Null Distribution Approach, Anal. Chem. 87 (2015) 3812–3819. https://doi.org/10.1021/ac504472s.

[39] L. Mikaliunaite, R.E. Synovec, Computational method for untargeted determination of cycling yeast metabolites using comprehensive two-dimensional gas chromatography time-of-flight mass spectrometry, Talanta 244 (2022) 123396. https://doi.org/10.1016/j.talanta.2022.123396.

[40] K. Murtada, D. Bowman, M. Edwards, J. Pawliszyn, Thin-film microextraction combined with comprehensive two-dimensional gas chromatography time-of-flight mass spectrometry screening for presence of multiclass organic pollutants in drinking water samples, Talanta 242 (2022) 123301. https://doi.org/10.1016/j.talanta.2022.123301.

[41] S. Schöneich, G.S. Ochoa, C.M. Monzón, R.E. Synovec, Minimum variance optimized Fisher ratio analysis of comprehensive two-dimensional gas chromatography / mass spectrometry data: Study of the pacu fish metabolome, J. Chromatogr. A 1667 (2022) 462868. https://doi.org/10.1016/j.chroma.2022.462868.

[42] P.E. Sudol, M. Galletta, P.Q. Tranchida, M. Zoccali, L. Mondello, R.E. Synovec, Untargeted profiling and differentiation of geographical variants of wine samples using headspace solid-phase microextraction flow-modulated comprehensive two-dimensional gas chromatography with the support of tile-based Fisher ratio analysis, J. Chromatogr. A 1662 (2022) 462735. https://doi.org/10.1016/j.chroma.2021.462735.

[43] Y. Zou, M. Gaida, F.A. Franchina, P. Stefanuto, J.-F. Focant, Distinguishing between Decaffeinated and Regular Coffee by HS-SPME-GC×GC-TOFMS, Chemometrics, and Machine Learning, Molecules 27 (2022) 1806. https://doi.org/10.3390/molecules27061806.

[44] C.N. Cain, T.J. Trinklein, G.S. Ochoa, R.E. Synovec, Tile-Based Pairwise Analysis of GC × GC-TOFMS Data to Facilitate Analyte Discovery and Mass Spectrum Purification, Anal. Chem. 94 (2022) 5658–5666. https://doi.org/10.1021/acs.analchem.2c00223.

[45] P.E. Sudol, G.S. Ochoa, C.N. Cain, R.E. Synovec, Tile-based variance rank initiated-unsupervised sample indexing for comprehensive two-dimensional gas chromatography-time-of-flight mass spectrometry, Anal. Chim. Acta 1209 (2022) 339847. https://doi.org/10.1016/j.aca.2022.339847.

[46] C.N. Cain, P.E. Sudol, K.L. Berrier, R.E. Synovec, Development of variance rank initiated-unsupervised sample indexing for gas chromatography-mass spectrometry analysis, Talanta 233 (2021) 122495. https://doi.org/10.1016/j.talanta.2021.122495.

[47] ASTM D445-19. Standard Test Method for Kinematic Viscosity of Transparent and Opaque Liquids (and Calculation of Dynamic Viscosity), West Conshohocken, PA, 2019. www.astm.org.

[48] ASTM D7171-16. Standard Test Method for Hydrogen Content of Middle Distillate Petroleum Products by Low-Resolution Pulsed Nuclear Magnetic Resonance Spectroscopy, West Conshohocken, PA, 2016. www.astm.org.

[49] ASTM D4809-18. Standard Test Method for Heat of Combustion of Liquid Hydrocarbon Fuels by Bomb Calorimeter (Precision Method), West Conshohocken, PA, 2018. www.astm.org.

[50] P.E. Sudol, G.S. Ochoa, R.E. Synovec, Investigation of the limit of discovery using tile-based Fisher ratio analysis with comprehensive two-dimensional gas chromatography time-of-flight mass spectrometry, J. Chromatogr. A 1644 (2021). https://doi.org/10.1016/j.chroma.2021.462092.

[51] C.N. Cain, S. Schöneich, R.E. Synovec, Development of an Enhanced Total Ion Current Chromatogram Algorithm to Improve Untargeted Peak Detection, Anal. Chem. 92 (2020) 11365–11373. https://doi.org/10.1021/acs.analchem.0c02136.

[52] J. Lee, J. Flores-Cerrillo, J. Wang, Q.P. He, A Variable Selection Method for Improving Variable Selection Consistency and Soft Sensor Performance, Proc. Am. Control Conf. 2020-July (2020) 725–730. https://doi.org/10.23919/ACC45564.2020.9147774.

[53] R. Bro, PARAFAC. Tutorial and applications, Chemom. Intell. Lab. Syst. 38 (1997) 149–171. https://doi.org/10.1016/S0169-7439(97)00032-4.

[54] R. Bro, Multiway calibration: Multilinear PLS, J. Chemom. 10 (1996) 47–61. https://doi.org/10.1002/(SICI)1099-128X(199601)10:1<47::AID-CEM400>3.0.CO;2-C.

[55] G. Knothe, K.R. Steidley, Kinematic viscosity of biodiesel fuel components and related compounds. Influence of compound structure and comparison to petrodiesel fuel components, Fuel 84 (2005) 1059–1065. https://doi.org/10.1016/j.fuel.2005.01.016.

[56] R.J. Urbanowicz, M. Meeker, W. La Cava, R.S. Olson, J.H. Moore, Relief-based feature selection: Introduction and review, J. Biomed. Inform. 85 (2018) 189–203. https://doi.org/10.1016/j.jbi.2018.07.014.

[57] S.H. Won, P.S. Veloo, S. Dooley, J. Santner, F.M. Haas, Y. Ju, F.L. Dryer, Predicting the global combustion behaviors of petroleum-derived and alternative jet fuels by simple fuel property measurements, Fuel 168 (2016) 34–46. https://doi.org/10.1016/j.fuel.2015.11.026.

[58] S.L. Outcalt, A. Laesecke, K.J. Brumback, Thermophysical properties measurements of rocket propellants RP-1 and RP-2, J. Propuls. Power 25 (2009) 1032–1040. https://doi.org/10.2514/1.40543.

[59] C. Yuan, D.A. Emelianov, M.A. Varfolomeev, M. Abaas, Comparison of oxidation behavior of linear and branched alkanes, Fuel Process. Technol. 188 (2019) 203–211. https://doi.org/10.1016/j.fuproc.2019.02.025.

Chapter 7: Tile-Based Pairwise Analysis of GC×GC-TOFMS Data to Facilitate Analyte Discovery and Mass Spectrum Purification

7.1. Introduction

The growing use of comprehensive two-dimensional (2D) gas chromatography coupled with time-of-flight mass spectrometry (GC×GC-TOFMS) in fields such as forensics [1–4], metabolomics [5–8], petroleomics [9–12], and food analysis [13–16] is due to the increased demand for a thorough characterization of the volatile and semi-volatile profile of these complex samples. While GC×GC-TOFMS is a powerful technique, manual interpretation of its information-rich data can be onerous. Chemometrics can be used to extract meaningful chemical information from the GC×GC-TOFMS data with minimal human intervention using both non-targeted and targeted methods.

Non-targeted analysis refers to the discovery of analytes in a data set that are responsible for the similarities/differences between samples. These chemometric methods can be labeled as supervised or unsupervised, where the former relies upon the knowledge of sample class membership. For example, tile-based Fisher ratio (F-ratio) analysis is a supervised, non-targeted technique that finds class distinguishing analytes in GC×GC-TOFMS data sets [17,18]. This method calculates the ratio of the between-class variance to the pooled within-class variance for the summed chromatographic signal within a small, rectangular section (i.e., tile) on a per-mass channel (*m/z*) basis [17,18]. Tile-based F-ratio analysis outputs a "hit list," which ranks analytes (i.e., "hits") in descending order of their F-ratios. Ideally, class distinguishing analytes will have

---

This chapter is reproduced from C. N. Cain, T. J. Trinklein, G. S. Ochoa, R. E. Synovec, Tile-Based Pairwise Analysis of GC×GC-TOFMS Data to Facilitate Analyte Discovery and Mass Spectrum Purification, Anal. Chem. 94 (2022), 5658-5666.

high F-ratios and land near the top of the hit list. Hence, multiple samples/replicates per class (e.g., 4 or 6 chromatograms per class) are necessary for this calculation. However, the time and materials required to collect replicate chromatograms may not always be available, especially for sample-limited or scouting studies. In these cases, an analyst would want to compare two GC×GC-TOFMS chromatograms to identify analytes that could potentially be different between them. Previous literature has proposed the use of subtraction plots [19–22] or Bayesian statistics [23] to compare two chromatograms. Subtraction plots are produced by direct subtraction of chromatograms belonging to different classes [19–22]. Whereas Bayesian methods calculate the Jensen-Shannon (JS) divergence on peak regions to find class distinguishing analytes [23]. Briefly, the JS divergence quantifies the similarity between the distributions of signal intensities within a tile on a per-*m/z* basis [23]. Since both methods are inherently pixel-based, they are susceptible to false positives (i.e., non-class distinguishing hits that are easily discovered) from either retention time misalignment, random detector fluctuations, and/or redundant hits [17].

Targeted chemometric methods can be used to identify and quantify the chromatographic peaks discovered by non-targeted analyses. Given the complexity of the samples separated by GC×GC-TOFMS, peaks discovered via non-targeted means may overlap with interferent signals. Decomposition methods like multivariate curve resolution-alternating least squares (MCR-ALS) [24] and parallel factor analysis (PARAFAC) [25] are often employed to mathematically resolve the pure chromatographic and *m/z* signals for overlapped peaks. However, the ability to obtain a high quality spectrum for analyte identification with MCR-ALS or PARAFAC is highly dependent on its 2D chromatographic resolution ($R_{s,2D}$) and extent of spectrum contamination, which can be measured as the ratio of the interferent signal to the target analyte signal (s_{int}/s_A) [26-31]. The poor performance of MCR-ALS can also be attributed to a type of uncertainty

known as rotational ambiguity, which describes the tendency of the model to return only one of many feasible solutions [32-34]. Recently, class comparison enabled-mass spectrum purification (CCE-MSP) has been introduced as an alternative to extract a pure spectrum for a target analyte [31]. The premise of CCE-MSP is that the target analyte spectrum changes between classes while the background spectrum is constant, although the extension to more than one target analyte was also demonstrated [31]. CCE-MSP obtains the purified target analyte spectrum by normalizing the hit spectra to a pure interferent *m/z* and then subtracting the two normalized spectra [31], which are identified using two signal consistency metrics, lack-of-fit (*LOF*) and *p*-value from a *t*-test [35]. For analytes discovered by F-ratio analysis, the pure analyte spectra produced with CCE-MSP had superior match values (MV) compared to the spectra obtained with MCR-ALS and PARAFAC [31].

To improve the discovery and identification of class distinguishing hits without requiring multiple samples/replicates, this report establishes a novel tile-based pairwise analysis method, termed 1v1 analysis. Tile-based 1v1 analysis utilizes the same four-grid tiling scheme developed for F-ratio analysis, which can reduce false positives while simultaneously increasing the signal-to-noise ratio (*S/N*) [17,18]. However, instead of calculating F-ratios, 1v1 analysis calculates a Rank Metric (*RM*), which is defined as the sum-normalized absolute difference,

$$RM = \frac{|s(m/z)_2 - s(m/z)_1|}{s(m/z)_2 + s(m/z)_1} \times 100 \qquad (7.1)$$

where $s(m/z)_1$ and $s(m/z)_2$ are the summed signals at a given *m/z* for a given tile in one sample (class 1) versus another sample (class 2). This calculation ensures that differences between peaks with both larger and smaller intensities can be readily discovered regardless of class assignment. Herein, tile-based 1v1 analysis is demonstrated on two complex data sets: a diesel fuel spiked with 18 non-native compounds at two different concentration levels and cacao beans affected by

moisture damage [36]. The resulting 1v1 hit lists are compared to the current standard non-targeted analyses (F-ratio analysis, subtraction plots, and JS divergence). This report also demonstrates that 1v1 analysis can be coupled to CCE-MSP for improvements in analyte identification. Furthermore, we propose an alternative decomposition method, termed CCE-MSP *assisted* MCR-ALS, which integrates these signal consistency metrics into MCR-ALS to resolve the targeted analytes discovered by 1v1 analysis. The pure analyte spectra obtained from these methods are then compared to standard MCR-ALS and PARAFAC. Ultimately, the workflow established herein can readily discover and identify analytes of interest using only one chromatogram per class.

7.2. Methods and Materials

Tile-based 1v1 analysis was performed on GC×GC-TOFMS separations of a diesel fuel spiked with 18 non-native compounds (Table E.1) and a previously investigated data set of cacao beans affected by moisture damage [36] (see Appendix E) using Matlab 2019b (Mathworks, Inc., Natick, MA, USA). Baseline drift in the chromatograms was corrected using a rolling ball minimum approach [37]. The spiked diesel data set was then normalized to the internal standard while the cacao bean data was normalized to the sum of the total ion current (TIC) chromatogram. For all paired analyses (tile-based 1v1 analysis, JS divergence, and pixel-based difference analysis), each replicate in one class was arbitrarily paired with a replicate in the other class, generating multiple hit lists. Meanwhile, F-ratio analysis was performed using all replicates in each class, generating one hit list. Subtraction plots for the pairwise comparisons were generated by computing the absolute difference between every pixel in the TIC or at every *m/z*. Meanwhile, 1v1, F-ratio, and JS divergence analyses were performed using the following conditions. A tile size of 16 s by 800 ms ($^1D \times {}^2D$) and cluster window size of 12 s × 750 ms was

selected for the spiked diesel data set while a tile size of 9 s × 300 ms and cluster window size of 7.5 s × 250 ms was chosen for the cacao bean data. These tile and cluster window sizes were selected based on the observed peak widths to prevent interferent signal from drowning out class distinguishing differences while minimizing the redundant hits in the final hit list [17,18,38]. A *S/N* threshold of 10 was used to eliminate tiles with insufficient signal [39]. Hit lists were ranked in descending order using the *m/z* that produced the maximum *RM*, F-ratio, or JS divergence [38].

The extension of 1v1 analysis to mass spectrum purification was demonstrated using the spiked diesel data set (Figure E.1). Chemometric decomposition was performed similarly to our previous publication [31] using PLS Toolbox 8.9 (Eigenvector Research, Inc., Wenatchee, WA, USA). Briefly, a 16 s × 800 ms section of data around each hit was isolated from each chromatogram. For MCR-ALS, the unfolded chromatograms were then augmented together to create a two-way array (retention × spectra). Likewise, PARAFAC was performed on the three-way array (retention × spectra × samples). The signal consistency metrics of *RM* and *LOF* were calculated for all *m/z* that passed the *S/N* threshold using the summed ^{2}D data within the tile [31,35]. Interferent *m/z* for each hit were defined as *m/z* with a $LOF \leq 5$ % and $RM \leq 5$ %. The pure analyte spectrum from CCE-MSP was obtained by subtracting the class 1 hit spectrum from the normalized class 2 hit spectrum, which is assigned to the chromatogram containing the more concentrated spikes (see Appendix E) [31]. CCE-MSP assisted MCR-ALS also obtained the pure analyte spectrum for the spiked analytes by performing MCR-ALS on the *m/z* not identified as pure interferent *m/z*. A forward MV was calculated by comparing the obtained analyte spectrum to the in-house library developed from the neat standards mixture. For comparison, a MV was

also calculated for the initial MCR-ALS results using the reduced set of *m/z* that was used for CCE-MSP assisted MCR-ALS.

7.3. Results and Discussion

The TIC chromatogram of the diesel fuel containing 18 spiked analytes is shown in Figure 7.1A. While this data set was devised to test the performance of our proposed workflow against standard non-targeted and targeted methods, the spiked analytes were chosen based on their low $R_{s,2D}$ and/or mass spectra similarity with compounds natively present in the diesel fuel. Hence, this represents a challenging data set for method validation. For the tile-based 1v1 analyses, replicates from both classes were paired up to discover class distinguishing analytes (i.e., the spiked analytes), generating six hit lists. With any non-targeted analysis, the feature ranking metric (e.g., *RM*) is calculated on any pixel, peak, or tile that is greater than the *S/N* threshold, causing the hit list to contain hits from both the class distinguishing analytes and the background matrix. The hit list is then analyzed from the top-down since hits near the top are more likely to be class distinguishing analytes. The distribution of the *RM* shown in Figure 7.1B demonstrates this concept, where the 18 spiked analytes (red bars) have the highest *RM*s (i.e., at the top of the hit list) while the native diesel analytes and redundant hits (gray bars) have smaller *RMs* (i.e., at the bottom of the hit list). Only one false positive due to a redundant hit was interspersed with the spiked analyte hits for the first 1v1 analysis results (Figure 7.1B). For the other five 1v1 analysis hit lists, all spiked analytes exhibited a *RM* of ~10 % or higher and were found in at least the top 24 hits, if not sooner (Table E.2). Note, since each analyte is spiked at a concentration ratio of 2, the ideal *RM* should be 33 % based on Eq 7.1. However, background signal from the native diesel compounds within the tile can cause deviations from the ideal (maximum) *RM*. Hits corresponding to native diesel components primarily had an $RM < 5$ %.

Thus, the *RM* easily differentiates between the spiked analytes and hits due to small variations from sample preparation and injection.

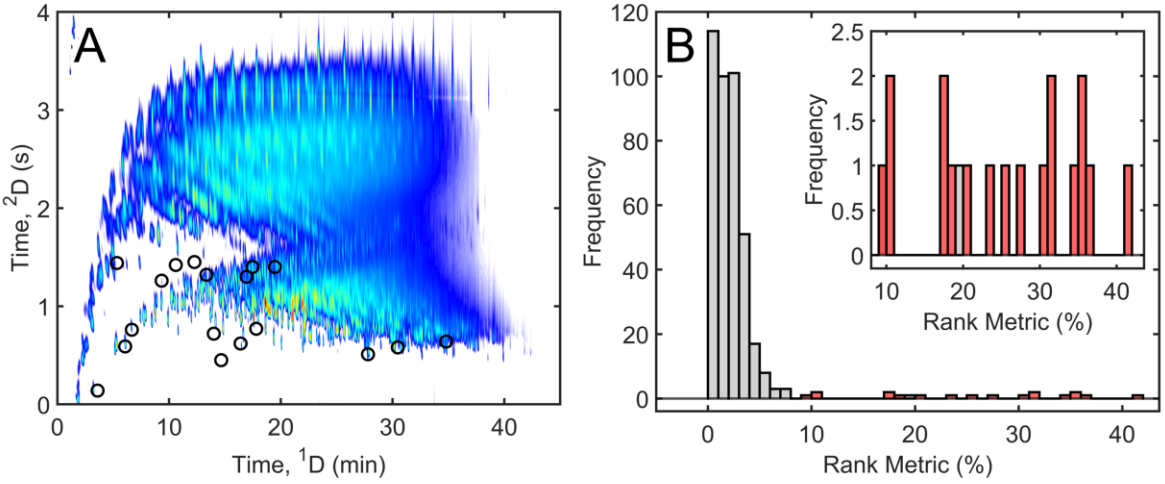


Figure 7.1. (A) Total ion current (TIC) chromatogram of the GC×GC-TOFMS separation of diesel fuel with circles indicating the locations of the 18 spiked analytes. (B) Distribution for the class comparison with tile-based 1v1 analysis using results for the first hit list. The spiked analytes are shown in red while all remaining hits are shown in gray.

Tile-based 1v1 analysis was compared to four standard methods for discovering class distinguishing analytes: tile-based F-ratio analysis [17,18], subtraction plots (using both the TIC and all *m/z*) [19–22], and JS divergence [23]. The hit lists for all these methods are provided in Table E.2-Table E.5. To objectively compare the results of the 1v1 analysis to these methods, receiver operating characteristic (ROC) curves were prepared for each non-targeted method. ROC curves can evaluate chemometric performance by illustrating the relationship between the true and false positive rates [39,40]. Figure 7.2 shows the ROC curves for each method using the top 100 hits in each list. The true positive rate was calculated as a running sum of true positives divided by the total number of true instances (i.e., 18 spiked analytes). Likewise, the false positive rate was calculated as a running sum of the number of false positives divided by the total number of false positives observed in the top 100 hits. Since multiple hit lists were produced for all paired comparisons (1v1 analysis, subtraction plots, and JS divergence), the rankings for each

analyte were averaged between the lists prior to generating the ROC curves in Figure 7.2. The area under the curve (AUC) for each ROC curve was also calculated (Figure 7.2) as a metric to compare the analyses. The AUC defines the probability that a randomly selected hit will be correctly categorized as either a true or false positive, where an AUC of 1 equals perfect ranking [39,40]. The AUCs for tile-based 1v1 (0.99) and F-ratio (0.98) analyses were the highest out of all the methods compared since these methods discovered all of the spiked analytes within the top 25 hits. Figure E.2 also shows the individual ROC curves for the six 1v1 hit lists, further highlighting their similar performance despite minor differences in their rankings. However, the AUCs for the JS divergence and subtraction plot methods ranged between 0.45 and 0.66, revealing the shortcomings of these methods. Namely, these methods are susceptible to false positives from retention time misalignment and spurious detector fluctuations since the data is not binned prior to their respective calculations. In turn, these false positives can obscure the discovery of true class-distinguishing hits. For example, 12 spiked analytes, on average, out of the total 18 were discovered in the top 100 hits using either the subtraction plots or JS divergence (Table E.3-Table E.5). Overall, Figure 7.2 illustrates that tile-based 1v1 analysis provides superior analyte discoverability compared to the other pairwise comparison methods.

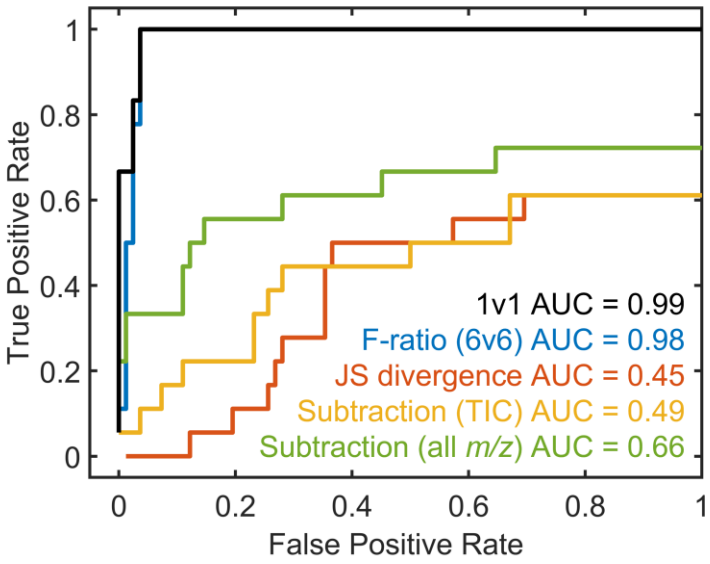


Figure 7.2. Receiver operating characteristic (ROC) curves for 1v1 analysis (black), F-ratio analysis (blue), Jensen-Shannon divergence (orange), subtraction plot using the TIC (yellow), and subtraction plot using all *m/z* (green). The respective area under the curve (AUC) is also provided.

To illustrate the challenge associated with the discovery of these true positives, an examination of methyl decanoate spiked at ~30 ppm is provided in Figure 7.3. As shown in the TIC chromatogram (Figure 7.3A), the chromatographic area surrounding this spiked analyte is highly overlapped. Figure 7.3B shows the same area at *m/z* 74, which had the highest *RM* in the 1v1 analyses. Since *m/z* 74 is free from background interference, the peak for methyl decanoate is clearly observed within the tile boundaries (dashed black box). Conversely, *m/z* 107 illustrates that a native diesel component is severely overlapped with methyl decanoate (Figure 7.3C). Using the retention times and peak widths in both dimensions [41], the $R_{s,2D}$ between methyl decanoate and the closest background interferent peak was 0.34. The s_{int}/s_A was also calculated, defined as the ratio between the summed signals of the interference spectrum divided by the summed signals from the pure analyte spectrum [31]. With a s_{int}/s_A of 29.8, the interferent signal from the native diesel background greatly overshadows the signal for methyl decanoate. This low

$R_{s,2D}$ and high s_{int}/s_A ultimately challenges analyte identification and quantitation. Indeed, Figure 7.3D compares the mass spectrum at the pin location discovered via tile-based 1v1 analysis (blue) and the in-house library spectrum (red). An initial MV of 159 was calculated for the two spectra, indicating the native diesel components contaminated the analyte mass spectrum. MCR-ALS and PARAFAC were performed on the area surrounding methyl decanoate to try to improve its MV. However, both methods failed to obtain a pure analyte spectrum with a MV ≥ 800, which is the standard for declaring a sufficient match [42], with MCR-ALS and PARAFAC producing a MV of 131 and 128, respectively (Table E.6). We note that optimizing separation conditions could improve the initial or chemometric resolved mass spectrum. However, this optimization is time consuming and impossible in a truly non-targeted study where it is not known which analytes change between classes. In contrast, α-pinene was easily identified with an average MV of 812 before chemometric decomposition because its signal was not swamped by the native diesel matrix (Table E.6; Figure E.3). Due to the low $R_{s,2D}$ and high s_{int}/s_A of many of the spiked analytes, CCE-MSP and CCE-MSP assisted MCR-ALS were applied to obtain a pure analyte spectrum for confident identification.

Recently, we demonstrated that our tile-based software can be extended to improve quantitation and analyte identification since the tiling procedure stores all the *m/z* discovered at each hit location [31,35]. As explained in our previous report [31] and Appendix E, CCE-MSP and CCE-MSP assisted MCR-ALS relies upon the identification of "sufficiently" pure interferent *m/z* in all the *m/z* stored by the tiling software. Two signal consistency metrics for 1v1 analysis were developed to find these interferent *m/z*: the *RM*, which quantitatively measures the signal differences between the two classes, and *LOF*, which determines the similarity in peak shapes between two classes.

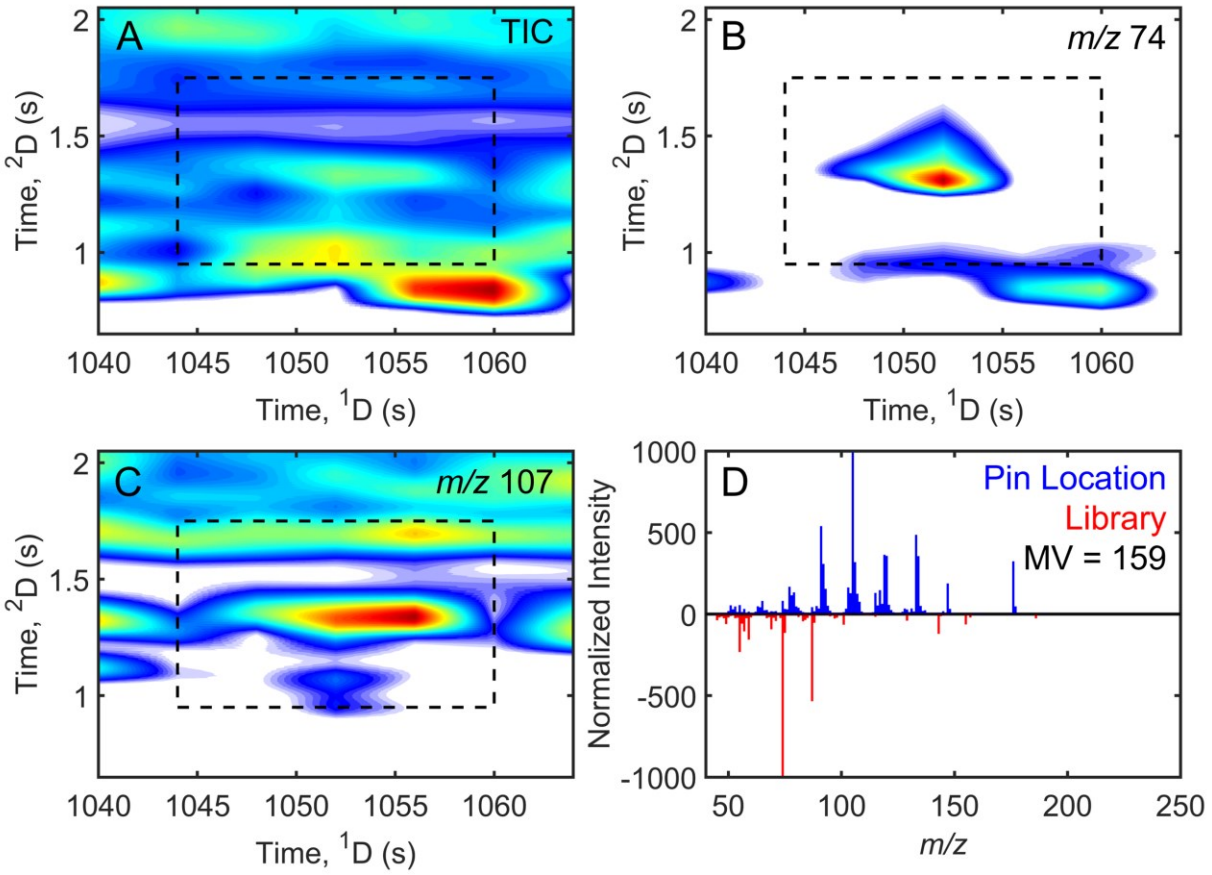


Figure 7.3. Illustration of the challenge identifying methyl decanoate, spiked at ~ 30 ppm, based on its match value (MV). (A) The 2D TIC chromatogram of the region around the analyte. The black dashed box represents the tile size of 16 s on ^{1}D and 800 ms on ^{2}D. (B) The 2D chromatogram at the top *m/z* discovered using 1v1 analysis. (C) The 2D chromatogram at an interferent *m/z*. (D) Comparison between the hit (blue) and library (red) spectra.

The signal consistency metrics developed for tile-based 1v1 analysis were applied to the discovered *m/z* for each hit. Plots of these signal consistency metrics are shown for methyl decanoate (Figure 7.4A) and all spiked analytes (Figure 7.4B). Ideally, *m/z* with a low *LOF* and *RM* are indicative of a relatively pure interferent *m/z* since the peak shape and signal does not change between classes. Strict *LOF* and *RM* thresholds were placed to ensure that the selected *m/z* had little to no signal contributions from the target analyte. Relatively pure interferent *m/z* for spectrum normalization of methyl decanoate were identified as *m/z* with a *LOF* and *RM* ≤ 5 % (Figure 7.4A; blue dots). Furthermore, previous work [31,35] determined that pure analyte *m/z*

can also be identified for quantitation efforts using plots similar to Figure 7.4A. For example, a pure analyte *m/z* for methyl decanoate (*m/z* 74) was identified as having a *LOF* ≤ 10 % and a *RM* ≥ 30 % (Figure 7.4A; green dot). Note, the same interferent and analyte *m/z* were also identified with F-ratio analysis (Figure E.4), validating the accuracy of the proposed signal consistency metrics for tile-based 1v1 analysis. Furthermore, the discovery of pure analyte *m/z* is easier with F-ratio analysis because the *t*-test provides statistical confidence [31,35]. Whereas tile-based 1v1 analysis does not provide this confidence since setting a *RM* threshold to discover pure analyte *m/z* for quantitation can be perilous if the concentration ratio is initially unknown. Therefore, for 1v1 analysis, it is recommended to use CCE-MSP assisted MCR-ALS (discussed later) to obtain the pure chromatographic peak profiles for quantitation. Figure 7.4B shows the signal consistency metrics for all 18 spiked analytes discovered with tile-based 1v1 analysis for the first paired chromatogram comparison. Similar plots corresponding to the other five 1v1 comparisons are shown in Figure E.5. In total, 191 *m/z* identified as belonging to the native diesel components and 39 *m/z* with potential contributions to analyte signal had a *RM* and *LOF* ≤ 5 % (Figure E.6). To further validate that the *LOF* and *RM* thresholds were not passing through *m/z* with class distinguishing differences, which could hinder mass spectrum purification, normalization factors were calculated for every *m/z* in Figure 7.4B with a *RM* and *LOF* ≤ 5 %. The normalization factors calculated for the interferent and potential analyte contributing *m/z* had averages of ~ 1 and were not statistically different (Figure E.7), confirming that the *m/z* below these thresholds effectively contain no class distinguishing signal.

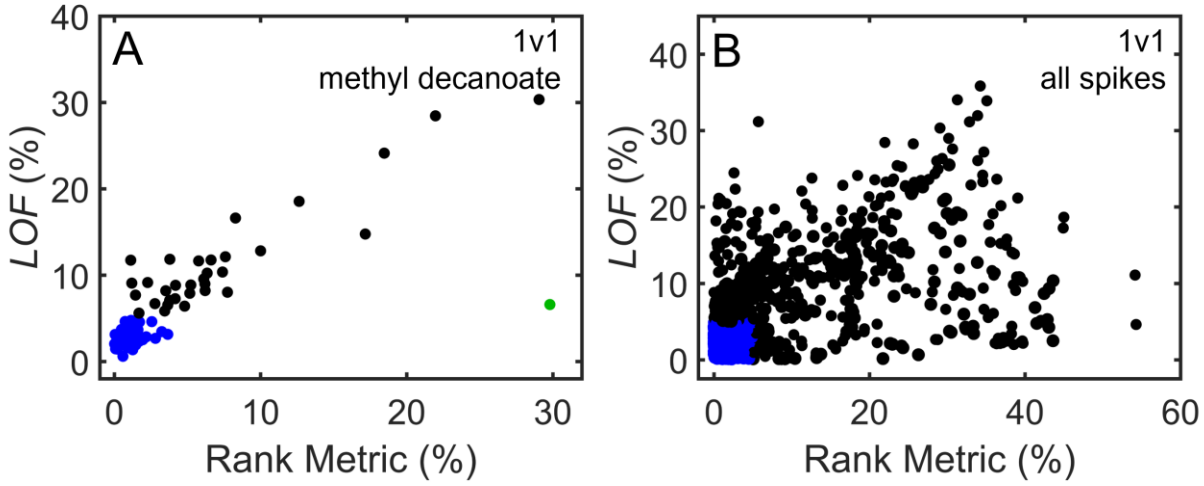


Figure 7.4. Illustration of the signal consistency metrics for 1v1 analysis for methyl decanoate (A) and all spiked analytes (B) using the first pairwise comparison. Pure interferent *m/z* (blue) were identified as *m/z* having a $LOF \leq 5$ % and a $RM \leq 5$ %. A pure analyte *m/z* for methyl decanoate is shown in green in (A).

Application of CCE-MSP to methyl decanoate using the pure interferent *m/z* discovered from tile-based 1v1 analysis is presented in Figure 7.5. The pure analyte *m/z* identified for methyl decanoate in Figure 7.4A (*m/z* 74; green bar) almost doubles in its intensity between class 1 and 2. Using this *m/z*, the signal ratio for methyl decanoate equals 1.88, which closely approximates the spiked concentration ratio of 2. Also, all interferent *m/z* identified in Figure 7.4A are highlighted by the blue bars in the hit spectra (Figure 7.5A-B), which demonstrates the degree of spectrum contamination by the interferences. For this example, the normalization factor was calculated using *m/z* 82, which is denoted by the blue star. The $S(m/z)_2$ shown in Figure 7.5A was normalized by the signal ratio for this interferent *m/z* ($k_1/k_2 = 1.01$). Since methyl decanoate had a s_{int}/s_A of 29.8, application of the secondary normalization ensures that the interference spectrum will not contaminate the analyte spectrum when the two hit spectra are subtracted [31]. The pure analyte spectrum for methyl decanoate (MV = 902) was obtained by subtracting $S(m/z)_1$ from the normalized $S(m/z)_2$ (Figure 7.5C). For all six 1v1 comparisons, an average MV of 908 was achieved with CCE-MSP, which is a significant improvement in

identification versus the initial hit spectrum (MV = 159) and resolved spectrum from MCR-ALS (MV = 131) and PARAFAC (MV = 128) (Table E.6). Although CCE-MSP is clearly beneficial to spectra highly contaminated by interferences, CCE-MSP also enhanced the identification of α-pinene from an initial MV of 812 to an average MV of 977 (Figure E.8). These results highlight the high-quality spectra enabled by tile-based 1v1 analysis.

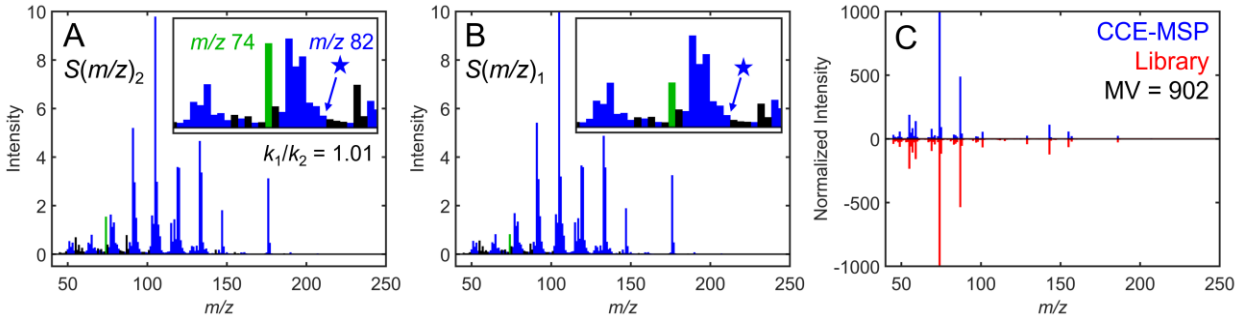


Figure 7.5. Application of CCE-MSP to methyl decanoate. The hit spectrum in class 2, $S(m/z)_2$ (A), and in class 1, $S(m/z)_1$ (B), are shown. The *m/z* shown in A and B are colored according to designation as pure interferent *m/z* (blue), pure analyte *m/z* (green), and all other *m/z* (black). $S(m/z)_2$ is normalized by k_1/k_2, which equates to the signal ratio for a pure interferent *m/z* (indicated by the blue star). Insets: Zoom-in from 60-90 *m/z* to illustrate the pure interferent and analyte *m/z*. The scale for the y-axes of the insets is -0.1 to 2. (C) Comparison of the purified analyte spectrum from CCE-MSP (blue) and the library spectrum (red). A match value (MV) is also provided.

CCE-MSP was performed on all 18 spiked analytes in the diesel fuel for all six 1v1 analyses for comparison to standard MCR-ALS and PARAFAC. Figure 7.6 demonstrates how the s_{int}/s_A and $R_{s,2D}$ affects analyte identification via initial MV at the pin location (A), MCR-ALS (B), PARAFAC (C), and CCE-MSP (D). Each data point represents one of the spiked analytes listed in Table E.6, colored according to the average MV determined for a given method. Figure 7.6A shows that only six analytes could initially be identified based on their mass spectrum at the hit location (MV ≥ 800). Use of MCR-ALS or PARAFAC was only able to increase the MVs for four more analytes above this threshold for identification (Figure 7.6B-C). In contrast, 17 out of the 18 spiked analytes were confidently identified with CCE-MSP

regardless of the $R_{s,2D}$ or s_{int}/s_A (Figure 7.6D). Note, while 2-dodecanone could not be identified by the MV ≥ 800 threshold after CCE-MSP, its MV was > 750, providing the analyst with an idea of possible compound classes or substituents. The inability to obtain high quality mass spectra via chemometric decomposition under challenging conditions like severe chromatographic overlap and large interference signals has been reported [26-31]. Rotational ambiguity can sometimes explain the poor performance of MCR-ALS [32-34]; however, the MCR-ALS results presented here do not appear affected by this type of uncertainty (Table E.7). Also, PARAFAC, which does produce unique solutions, was unable to produce a high-quality mass spectrum for many of the analytes despite the trilinear behavior of the data set (Table E.6). Therefore, we believe the inadequate performance of these decomposition methods is due to a "chemometric" multiplex disadvantage [31].

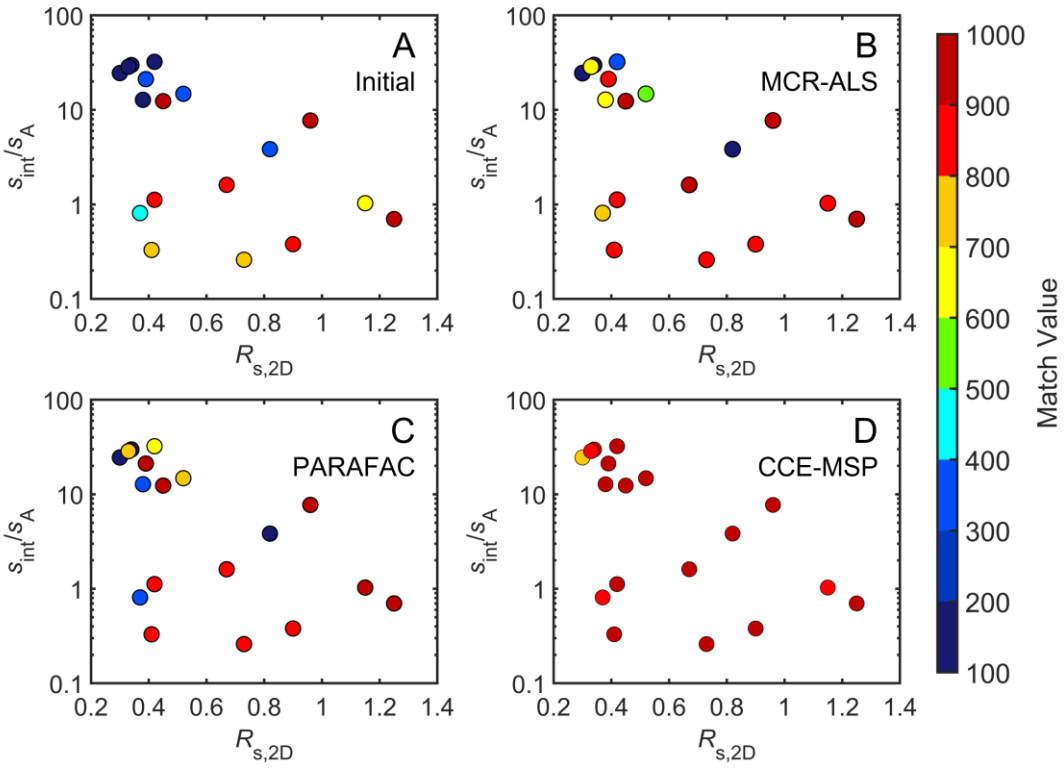


Figure 7.6. Plots of the interference-to-analyte ratios (s_{int}/s_A) versus 2D resolution ($R_{s,2D}$) for the 18 spiked analytes. Each point is colored according to the average MV determined initially (A) or with MCR-ALS (B), PARAFAC (C), and CCE-MSP (D).

Conceptually, the chemometric multiplex disadvantage is analogous to the instrumental multiplex disadvantage recognized for various forms of spectrometry, for instrumental designs where the simultaneous measurement of smaller signals at a given wavelength is hindered by noise from larger signals at all other wavelengths [43]. Similar to these simultaneous collection spectrometers, these chemometric methods aim to simultaneously model all the components present to minimize the residual error. As a result, there is "computational cross-talk" across the *m/z* domain, and noise and signal from the larger interferents can be included into the decomposed mass spectrum for the target analyte. This disadvantage cannot be disentangled from the traditional higher-order advantages obtained by coupling multidimensional instruments with chemometrics [44]. CCE-MSP overcomes this disadvantage by using these large interferent signals "against themselves" as a way to normalize the two mass spectra before their subtraction.

Further evidence of the chemometric multiplex disadvantage was uncovered. Generally, MCR-ALS decomposition is performed using all *m/z* collected to resolve all peaks in the region of interest. When MCR-ALS was initially performed using all *m/z* collected, it was unable to resolve methyl decanoate because the model was dominated by its effort to resolve the larger background interferences (Figure E.9). Thus, the chemometric multiplex disadvantage occurs. To overcome this disadvantage, MCR-ALS was performed on the spiked analytes using only the *m/z* not identified as interferent *m/z* based on the *RM* and *LOF* thresholds. We refer to this method as CCE-MSP *assisted* MCR-ALS. A total of 33 *m/z* discovered for methyl decanoate (albeit with some interferent contribution) had *RM* and *LOF* $\geq$ 5 % (Figure 7.4A). Visual inspection of the analytical ion current (AIC) chromatogram created by summing those 33 *m/z* (Figure 7.7A) reveals that there is signal changing between classes. Therefore, CCE-MSP assisted MCR-ALS operates by reducing the data input to MCR-ALS down to these 33 *m/z* identified as having a *RM* and *LOF*

≥ 5 %. Figure 7.7B-C shows that CCE-MSP assisted MCR-ALS was able to resolve both the pure chromatographic peak profiles and mass spectrum (MV = 967) for methyl decanoate. Note, analyte quantitation can be readily performed since CCE-MSP assisted MCR-ALS produces the pure chromatographic peak profiles. To compare the initial MCR-ALS results to the CCE-MSP assisted MCR-ALS, the initial MCR-ALS mass spectrum obtained in Figure E.9 was filtered down to the same 33 *m/z* of interest and a MV to the filtered library spectrum was calculated (Figure 7.7D). With a MV of 605, Figure 7.7D shows that some signal from the larger interferents was still included in these *m/z*, providing evidence of "computational cross-talk". Application of CCE-MSP assisted MCR-ALS to α-pinene also showed major improvement in the quality of the resolved chromatographic profiles and spectrum (Figure E.10). Overall, 17 of the 18 spiked analytes could be confidently identified with MV > 800 using CCE-MSP assisted MCR-ALS regardless of their s_{int}/s_A and $R_{s,2D}$ (Figure E.11). This performance is like applying the original CCE-MSP method on the 1v1 outputs. CCE-MSP assisted MCR-ALS was able to better capture the signal from these highly interfered spiked analytes because a majority of the noisy and/or interferent *m/z* were removed by the thresholding step shown in Figure 7.4, suppressing the chemometric multiplex disadvantage. Figure E.11 also compares the MV of the initial MCR-ALS spectrum filtered down to the analytes of interest. Except for a couple analytes that were well decomposed by MCR-ALS using all *m/z*, the MCR-ALS spectrum for many of the other analytes had some degree of contamination from this "computational cross-talk" since these MVs were less than the MVs obtained by CCE-MSP assisted MCR-ALS.

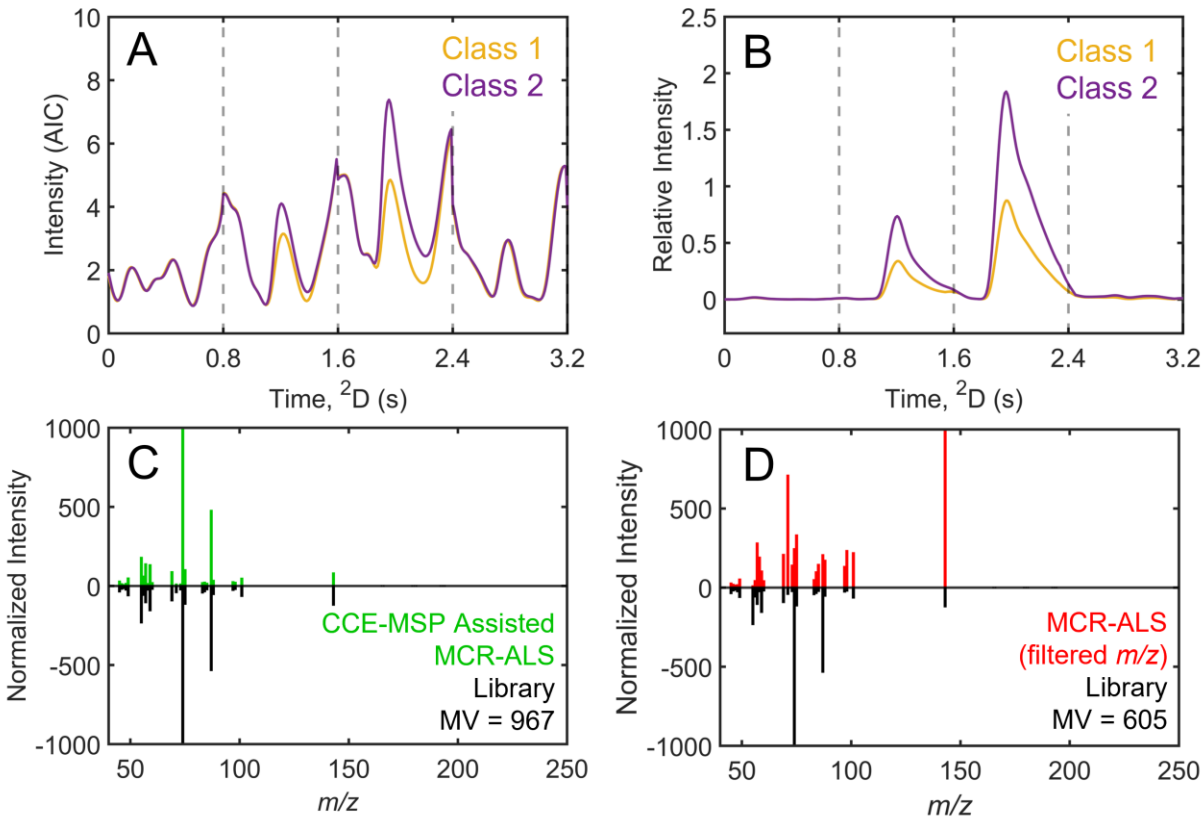


Figure 7.7. Demonstration of CCE-MSP assisted MCR-ALS on methyl decanoate. (A) The unfolded analytical ion current (AIC) chromatograms for class 1 (yellow) and class 2 (purple). The AICs represent the sum of the 33 *m/z* above the *RM* and *LOF* thresholds. The gray dashed vertical lines represent each modulation in the tile. (B) The chromatographic peak profiles for the CCE-MSP assisted MCR-ALS component that had the highest MV to the library spectrum. (C) Comparison of the mass spectrum for the CCE-MSP assisted MCR-ALS component in B (green) to the filtered library spectrum of methyl decanoate (black). (D) Reflection plot of the initial standard MCR-ALS spectrum, filtered down to the 33 *m/z* of interest (red), and the filtered library spectrum (black).

Non-targeted chemometric methods have been beneficial in classifying food samples based on their sensory profile or botanical and geographic origin as well as in monitoring biological changes due to environmental processes [45,46]. Herein, tile-based 1v1 analysis was applied to discover analytes indicative of moisture damage in cacao beans that presents both food safety and quality concerns [36]. Figure 7.8 illustrates that 1v1 analysis discovered 86 peaks with at least a 2-fold concentration change between the unmolded (A) and molded (B) chromatograms. This magnitude of concentration change is generally accepted to be indicative of

true chemical differences rather than natural variation [47]. Of those analytes identified (Table E.8), 72 of them had a higher signal in the unmolded beans (indicated by black rectangles on Figure 7.8) and 14 had a higher signal in molded beans (indicated by red rectangles on Figure 7.8). Many of these analytes were also identified in the top 30 hits of the other 1v1 and F-ratio hit lists (Table E.9-Table E.13), consistent with previous studies investigating cacao quality [36,48]. For example, analytes that had a higher abundance in the unmolded beans like 2,3-butanediol, butanoic acid, 2-ethyl-1-hexanol, and nonanal have been described to have sweet, creamy, and citrus-like aromas [48]. Conversely, analytes that had higher abundance in the molded beans like trimethylpyrazine, tetramethylpyrazine, 2,3-dimentyl pyrazine and 1-octen-3-ol contribute earthy and moldy aromas to the cacao beans [48]. Ultimately, 1v1 analysis was capable of discovering biomarkers of moisture damage in cacao beans, demonstrating that this technique can be routinely applied for quality control studies.

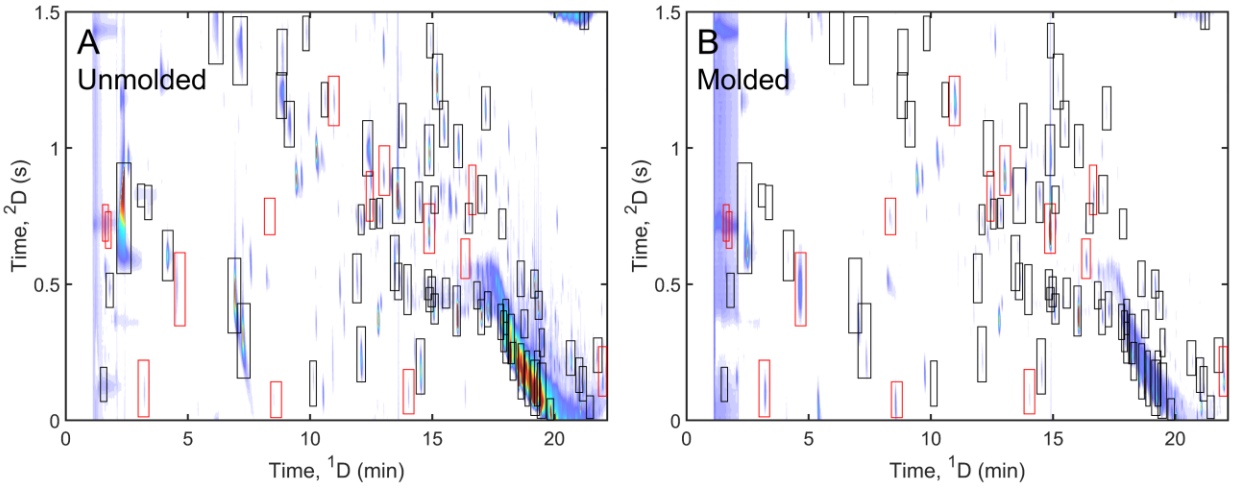


Figure 7.8. TIC chromatograms of the unmolded (A) and molded (B) cacao beans. The black rectangles indicate peaks that were higher in the unmolded sample and red rectangles indicate peaks that were higher in the molded sample.

7.4. Conclusion

Tile-based 1v1 analysis is a powerful supervised method for discovering meaningful chemical differences between two chromatograms. This methodology was able to discover the 18 analytes that were spiked into diesel fuel at concentrations as low as 10 ppm and in highly saturated chromatographic areas. Performance of tile-based 1v1 analysis was essentially equivalent to tile-based F-ratio analysis for pairwise comparisons as illustrated with ROC curves, and superior to pairwise comparisons based upon either chromatogram subtraction or JS divergence methods. This report also established a new signal consistency metric, the *RM*, to discover pure interferent *m/z* in conjunction with the *LOF* metric. These pure interferent *m/z* were necessary to obtain high quality spectra for analyte identification. The MVs for the spectra produced via CCE-MSP and CCE-MSP assisted MCR-ALS were far superior to the spectra generated with common chemometric decomposition methods like standard MCR-ALS and PARAFAC, especially for analytes with a $R_{s,2D} < 0.5$ and/or a $s_{int}/s_A > 10$. These results provide evidence of a chemometric multiplex disadvantage [31]. Furthermore, tile-based 1v1 analysis was capable of discovering biomarkers indicative of moisture damage in cacao beans. Overall, the data analysis workflow established herein can be utilized with any gas or liquid chromatographic data sets to discover and identify potential analytes of interest for sample-limited or scouting studies, where the collection of multiple replicates is unavailable. Future work aims to demonstrate the capabilities of tile-based 1v1 analysis to discover pertinent analytes in a variety of chromatographic applications.

7.5. References

[1] P.H. Stefanuto, K.A. Perrault, S. Stadler, R. Pesesse, H.N. Leblanc, S.L. Forbes, J.F. Focant, GC×GC-TOFMS and supervised multivariate approaches to study human cadaveric decomposition olfactive signatures, Anal. Bioanal. Chem. 407 (2015) 4767–4778. https://doi.org/10.1007/s00216-015-8683-5.

[2] L.M. Dubois, K.A. Perrault, P.-H. Stefanuto, S. Koschinski, M. Edwards, L. McGregor, J.-F. Focant, Thermal desorption comprehensive two-dimensional gas chromatography coupled to variable-energy electron ionization time-of-flight mass spectrometry for monitoring subtle changes in volatile organic compound profiles of human blood, J. Chromatogr. A. 1501 (2017) 117–127. https://doi.org/10.1016/j.chroma.2017.04.026.

[3] M. Lopatka, A.A. Sampat, S. Jonkers, L.A. Adutwum, H.G.J. Mol, G. van der Weg, J.J. Harynuk, P.J. Schoenmakers, A. van Asten, M.J. Sjerps, G. Vivó-Truyols, Local Ion Signatures (LIS) for the examination of comprehensive two-dimensional gas chromatography applied to fire debris analysis, Forensic Chem. 3 (2017) 1–13. https://doi.org/10.1016/j.forc.2016.10.003.

[4] A.A.S. Sampat, B. Van Daelen, M. Lopatka, H. Mol, G. Van Derweg, G.V. Truyols, M. Sjerps, P.J. Schoenmakers, A.C.V. Asten, Detection and characterization of ignitable liquid residues in forensic fire debris samples by comprehensive two-dimensional gas chromatography, Separations. 5 (2018) 1–27. https://doi.org/10.3390/separations5030043.

[5] H.D. Bean, J.E. Hill, J.M.D. Dimandja, Improving the quality of biomarker candidates in untargeted metabolomics via peak table-based alignment of comprehensive two-dimensional gas chromatography-mass spectrometry data, J. Chromatogr. A. 1394 (2015) 111–117. https://doi.org/10.1016/j.chroma.2015.03.001.

[6] G. Purcaro, P.H. Stefanuto, F.A. Franchina, M. Beccaria, W.F. Wieland-Alter, P.F. Wright, J.E. Hill, SPME-GC×GC-TOFMS fingerprint of virally-infected cell culture: Sample preparation optimization and data processing evaluation, Anal. Chim. Acta. 1027 (2018) 158–167. https://doi.org/10.1016/j.aca.2018.03.037.

[7] N. Di Giovanni, M.A. Meuwis, E. Louis, J.F. Focant, Untargeted Serum Metabolic Profiling by Comprehensive Two-Dimensional Gas Chromatography-High-Resolution Time-of-Flight Mass Spectrometry, J. Proteome Res. 19 (2020) 1013–1028. https://doi.org/10.1021/acs.jproteome.9b00535.

[8] M. Cialiè Rosso, F. Stilo, S. Squara, E. Liberto, S. Mai, C. Mele, P. Marzullo, G. Aimaretti, S.E. Reichenbach, M. Collino, C. Bicchi, C. Cordero, Exploring extra dimensions to capture saliva metabolite fingerprints from metabolically healthy and unhealthy obese patients by comprehensive two-dimensional gas chromatography featuring Tandem Ionization mass spectrometry, Anal. Bioanal. Chem. (2020). https://doi.org/10.1007/s00216-020-03008-6.

[9] R.L. Webster, P.M. Rawson, C. Kulsing, D.J. Evans, P.J. Marriott, Investigation of the Thermal Oxidation of Conventional and Alternate Aviation Fuels with Comprehensive Two-Dimensional Gas Chromatography Accurate Mass Quadrupole Time-of-Flight Mass Spectrometry, Energy Fuels. 31 (2017) 4886–4894. https://doi.org/10.1021/acs.energyfuels.7b00178.

[10] L. Bai, J. Smuts, J. Schenk, J. Cochran, K.A. Schug, Comparison of GC-VUV, GC-FID, and comprehensive two-dimensional GC–MS for the characterization of weathered and unweathered diesel fuels, Fuel. 214 (2018) 521–527. https://doi.org/10.1016/j.fuel.2017.11.053.

[11] G.L. Alexandrino, J. Malmborg, F. Augusto, J.H. Christensen, Investigating weathering in light diesel oils using comprehensive two-dimensional gas chromatography–High resolution mass spectrometry and pixel-based analysis: Possibilities and limitations, J. Chromatogr. A. 1591 (2019) 155–161. https://doi.org/10.1016/j.chroma.2019.01.042.

[12] A. Moreira de Oliveira, C. Alberto Teixeira, L. Wang Hantao, Evaluation of the retention profile in flow-modulated comprehensive two-dimensional gas chromatography and independent component analysis of weathered heavy oils, Microchem. J. 172 (2022) 106978. https://doi.org/10.1016/j.microc.2021.106978.

[13] D.D. Yan, Y.F. Wong, L. Tedone, R.A. Shellie, P.J. Marriott, S.P. Whittock, A. Koutoulis, Chemotyping of new hop (Humulus lupulus L.) genotypes using comprehensive two-dimensional gas chromatography with quadrupole accurate mass time-of-flight mass spectrometry, J. Chromatogr. A. 1536 (2018) 110–121. https://doi.org/10.1016/j.chroma.2017.08.020.

[14] C. Cordero, A. Guglielmetti, C. Bicchi, E. Liberto, L. Baroux, P. Merle, Q. Tao, S.E. Reichenbach, Comprehensive two-dimensional gas chromatography coupled with time of flight mass spectrometry featuring tandem ionization: Challenges and opportunities for accurate fingerprinting studies, J. Chromatogr. A. 1597 (2019) 132–141. https://doi.org/10.1016/j.chroma.2019.03.025.

[15] F.J.M. Novaes, A.I. da Silva Junior, C. Kulsing, Y. Nolvachai, H.R. Bizzo, F.R. de Aquino Neto, C.M. Rezende, P.J. Marriott, New approaches to monitor semi-volatile organic compounds released during coffee roasting using flow-through/active sampling and comprehensive two-dimensional gas chromatography, Food Res. Int. 119 (2019) 349–358. https://doi.org/10.1016/j.foodres.2019.02.009.

[16] P.E. Sudol, M. Galletta, P.Q. Tranchida, M. Zoccali, L. Mondello, R.E. Synovec, Untargeted profiling and differentiation of geographical variants of wine samples using headspace solid-phase microextraction flow-modulated comprehensive two-dimensional gas chromatography with the support of tile-based Fisher ratio analysis, J. Chromatogr. A. 1662 (2022) 462735. https://doi.org/10.1016/j.chroma.2021.462735.

[17] L.C. Marney, W.C. Siegler, B.A. Parsons, J.C. Hoggard, B.W. Wright, R.E. Synovec, Tile-based Fisher-ratio software for improved feature selection analysis of comprehensive two-dimensional gas chromatography-time-of-flight mass spectrometry data, Talanta. 115 (2013) 887–895. https://doi.org/10.1016/j.talanta.2013.06.038.

[18] B.A. Parsons, L.C. Marney, W.C. Siegler, J.C. Hoggard, B.W. Wright, R.E. Synovec, Tile-Based Fisher Ratio Analysis of Comprehensive Two-Dimensional Gas Chromatography Time-of-Flight Mass Spectrometry (GC × GC–TOFMS) Data Using a Null Distribution Approach, Anal. Chem. 87 (2015) 3812–3819. https://doi.org/10.1021/ac504472s.

[19] R.A. Shellie, W. Welthagen, J. Zrostliková, J. Spranger, M. Ristow, O. Fiehn, R. Zimmermann, Statistical methods for comparing comprehensive two-dimensional gas chromatography-time-of-flight mass spectrometry results: Metabolomic analysis of mouse tissue extracts, J. Chromatogr. A. 1086 (2005) 83–90. https://doi.org/10.1016/j.chroma.2005.05.088.

[20] T.C. Tran, P.J. Marriott, Characterization of incense smoke by solid phase microextraction-Comprehensive two-dimensional gas chromatography (GC×GC), Atmos. Environ. 41 (2007) 5756–5768. https://doi.org/10.1016/j.atmosenv.2007.02.030.

[21] M. Kallio, M. Kivilompolo, S. Varjo, M. Jussila, T. Hyötyläinen, Data analysis programs for comprehensive two-dimensional chromatography, J. Chromatogr. A. 1216 (2009) 2923–2927. https://doi.org/10.1016/j.chroma.2008.11.037.

[22] G.T. Ventura, B.R.T. Simoneit, R.K. Nelson, C.M. Reddy, The composition, origin and fate of complex mixtures in the maltene fractions of hydrothermal petroleum assessed by comprehensive two-dimensional gas chromatography, Org. Geochem. 45 (2012) 48–65. https://doi.org/10.1016/j.orggeochem.2012.01.002.

[23] A. Barcaru, G. Vivó-Truyols, Use of Bayesian Statistics for Pairwise Comparison of Megavariate Data Sets: Extracting Meaningful Differences between GCxGC-MS Chromatograms Using Jensen–Shannon Divergence, Anal. Chem. 88 (2016) 2096–2104. https://doi.org/10.1021/acs.analchem.5b03506.

[24] R. Tauler, Multivariate curve resolution applied to second order data, Chemom. Intell. Lab. Syst. 30 (1995) 133–146. https://doi.org/10.1016/0169-7439(95)00047-X.

[25] R. Bro, PARAFAC. Tutorial and applications, Chemom. Intell. Lab. Syst. 38 (1997) 149–171. https://doi.org/10.1016/S0169-7439(97)00032-4.

[26] I.H.M. van Stokkum, K.M. Mullen, V. V. Mihaleva, Global analysis of multiple gas chromatography-mass spectrometry (GC/MS) data sets: A method for resolution of co-eluting components with comparison to MCR-ALS, Chemom. Intell. Lab. Syst. 95 (2009) 150–163. https://doi.org/10.1016/j.chemolab.2008.10.004.

[27] H.P. Bailey, S.C. Rutan, P.W. Carr, Factors that affect quantification of diode array data in comprehensive two-dimensional liquid chromatography using chemometric data analysis, J. Chromatogr. A. 1218 (2011) 8411–8422. https://doi.org/10.1016/j.chroma.2011.09.057.

[28] X. Domingo-Almenara, A. Perera, N. Ramírez, N. Cañellas, X. Correig, J. Brezmes, Compound identification in gas chromatography/mass spectrometry-based metabolomics by blind source separation, J. Chromatogr. A. 1409 (2015) 226–233. https://doi.org/10.1016/j.chroma.2015.07.044.

[29] A. Eftekhari, H. Parastar, Multivariate analytical figures of merit as a metric for evaluation of quantitative measurements using comprehensive two-dimensional gas chromatography–mass spectrometry, J. Chromatogr. A. 1466 (2016) 155–165. https://doi.org/10.1016/j.chroma.2016.09.016.

[30] D.K. Pinkerton, B.C. Reaser, K.L. Berrier, R.E. Synovec, Determining the probability of achieving a successful quantitative analysis for gas chromatography-mass spectrometry, Anal. Chem. 89 (2017) 9926–9933. https://doi.org/10.1021/acs.analchem.7b02230.

[31] G.S. Ochoa, P.E. Sudol, T.J. Trinklein, R.E. Synovec, Class comparison enabled mass spectrum purification for comprehensive two-dimensional gas chromatography with time-of-flight mass spectrometry, Talanta. 236 (2022) 122844. https://doi.org/10.1016/j.talanta.2021.122844.

[32] J. Jaumot, R. Tauler, MCR-BANDS: A user friendly MATLAB program for the evaluation of rotation ambiguities in Multivariate Curve Resolution, Chemom. Intell. Lab. Syst. 103 (2010) 96–107. https://doi.org/10.1016/j.chemolab.2010.05.020.

[33] R.B. Pellegrino Vidal, A.C. Olivieri, R. Tauler, Quantifying the Prediction Error in Analytical Multivariate Curve Resolution Studies of Multicomponent Systems, Anal. Chem. 90 (2018) 7040–7047. https://doi.org/10.1021/acs.analchem.8b01431.

[34] X. Zhang, Z. Zhang, R. Tauler, Evaluation of the extension of rotation ambiguity associated to multivariate curve resolution solutions by the application of the MCR-BANDS method, Talanta. 202 (2019) 554–564. https://doi.org/10.1016/j.talanta.2019.05.002.

[35] G.S. Ochoa, S.E. Prebihalo, B.C. Reaser, L.C. Marney, R.E. Synovec, Statistical inference of mass channel purity from Fisher ratio analysis using comprehensive two-dimensional gas chromatography with time of flight mass spectrometry data, J. Chromatogr. A. 1627 (2020) 461401. https://doi.org/10.1016/j.chroma.2020.461401.

[36] E.M. Humston, J.D. Knowles, A. McShea, R.E. Synovec, Quantitative assessment of moisture damage for cacao bean quality using two-dimensional gas chromatography combined with time-of-flight mass spectrometry and chemometrics, J. Chromatogr. A. 1217 (2010) 1963–1970. https://doi.org/10.1016/j.chroma.2010.01.069.

[37] H.G. Schmarr, J. Bernhardt, Profiling analysis of volatile compounds from fruits using comprehensive two-dimensional gas chromatography and image processing techniques, J. Chromatogr. A. 1217 (2010) 565–574. https://doi.org/10.1016/j.chroma.2009.11.063.

[38] P.E. Sudol, G.S. Ochoa, R.E. Synovec, Investigation of the limit of discovery using tile-based Fisher ratio analysis with comprehensive two-dimensional gas chromatography time-of-flight mass spectrometry, J. Chromatogr. A. 1644 (2021). https://doi.org/10.1016/j.chroma.2021.462092.

[39] B.C. Reaser, B.W. Wright, R.E. Synovec, Using Receiver Operating Characteristic Curves to Optimize Discovery-Based Software with Comprehensive Two-Dimensional Gas Chromatography with Time-of-Flight Mass Spectrometry, Anal. Chem. 89 (2017) 3606–3612. https://doi.org/10.1021/acs.analchem.6b04991.

[40] C.D. Brown, H.T. Davis, Receiver operating characteristics curves and related decision measures: A tutorial, Chemom. Intell. Lab. Syst. 80 (2006) 24–38. https://doi.org/10.1016/j.chemolab.2005.05.004.

[41] R.E. Murphy, M.R. Schure, J.P. Foley, Effect of Sampling Rate on Resolution in Comprehensive Two-Dimensional Liquid Chromatography, Anal. Chem. 70 (1998) 1585–1594. https://doi.org/10.1021/ac971184b.

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

[43] E. Voigtman, J.D. Winefordner, The Multiplex Disadvantage and Excess Low-Frequency Noise, Appl. Spectrosc. 41 (1987) 1182–1184. https://doi.org/10.1366/0003702874447509.

[44] K.S. Booksh, B.R. Kowalski, Theory of Analytical Chemistry, Anal. Chem. 66 (1994) 782–791. https://doi.org/10.1021/ac00087a001.

[45] C. Cordero, J. Kiefl, P. Schieberle, S.E. Reichenbach, C. Bicchi, Comprehensive two-dimensional gas chromatography and food sensory properties: Potential and challenges, Anal. Bioanal. Chem. 407 (2015) 169–191. https://doi.org/10.1007/s00216-014-8248-z.

[46] F. Stilo, C. Bicchi, A. Robbat, S.E. Reichenbach, C. Cordero, Untargeted approaches in food-omics: The potential of comprehensive two-dimensional gas chromatography/mass spectrometry, TrAC - Trends Anal. Chem. 135 (2021) 116162. https://doi.org/10.1016/j.trac.2020.116162.

[47] K. Ortmayr, V. Charwat, C. Kasper, S. Hann, G. Koellensperger, Uncertainty budgeting in fold change determination and implications for non-targeted metabolomics studies in model systems, Analyst. 142 (2017) 80–90. https://doi.org/10.1039/c6an01342b.

[48] F. Magagna, A. Guglielmetti, E. Liberto, S.E. Reichenbach, E. Allegrucci, G. Gobino, C. Bicchi, C. Cordero, Comprehensive Chemical Fingerprinting of High-Quality Cocoa at Early Stages of Processing: Effectiveness of Combined Untargeted and Targeted Approaches for Classification and Discrimination, J. Agric. Food Chem. 65 (2017) 6329–6341. https://doi.org/10.1021/acs.jafc.7b02167.

Chapter 8: Development of an Enhanced Total Ion Current Chromatogram Algorithm to Improve Untargeted Peak Detection

8.1. Introduction

Demands for increased resolution of complex samples containing volatile and semi-volatile analytes, and/or those amenable to gas phase analysis has led to the wide implementation of comprehensive two-dimensional (2D) gas chromatography (GC×GC). This powerful separation technique was introduced in the early 1990's, wherein a modulator continuously collects and injects effluent from the first dimension column (^{1}D) onto the second dimension (^{2}D) column [1]. Since its introduction, a variety of modulation devices have been developed in order to both reduce separation times, generate narrower peak widths, and increase peak capacity [2]. GC×GC has been shown to increase the peak capacity of a separation by an order of magnitude compared to a one-dimensional GC (1D-GC) separation with constant run times [3]. Furthermore, the ^{1}D and ^{2}D separation mechanisms should be complementary for analytes to have different interactions with the stationary phase, increasing the selectivity and resolution of the separation [4]. Therefore, GC×GC has been applied to a variety of complex matrices in fields ranging from metabolomics [5–8], environmental monitoring [9–11], flavor and fragrance monitoring [12–14], forensics [15–17], and fuel analysis [18,19]. The use of GC×GC with multichannel detection, such as time-of-flight mass spectrometry (TOFMS), often requires advanced chemometric techniques to extract meaningful chemical information from the large quantity of data produced [20].

Optimized chemometric performance relies on the use of pre-processing methods, which

---

This chapter is reproduced from C. N. Cain, S. Schöneich, R. E. Synovec, Development of an Enhanced Total Ion Current Chromatogram Algorithm to Improve Untargeted Peak Detection, Anal. Chem. 92 (2020), 11365-11373.

are categorized into four categories: data reduction, peak alignment, noise and baseline correction, and peak detection [21,22]. Perhaps the most important pre-processing step, regardless of the analytical objective, is peak detection (*i.e.*, the ability to distinguish analyte signals from a noisy background). Peak detection, for either targeted or nontargeted approaches, is primarily based on distinguishing an analyte signal (*i.e.*, Gaussian-like peak) from the background noise. Most of the current peak detection methods for 2D chromatography rely upon the adaptation of methods developed for 1D chromatography [22]. A two-step peak detection method was originally proposed, where the first derivative is used to detect peaks on the ^{1}D and a decision tree is used to cluster those detected peaks in the ^{2}D [23]. An alternative method for 2D peak detection relies upon image analysis techniques, such as watershed-based algorithms [24]. Watershed-based algorithms detect the maxima in a 2D image chromatogram and evaluate the signal of neighboring pixels to determine if those signals contribute to the chemical information or background noise [25,26]. Statistical approaches using Bayesian inference [27,28] and probability models [29,30] have also been explored to identify chromatographic regions with high likelihood of containing peaks.

These peak detection algorithms are standardly performed using the total ion current chromatogram (TIC), which is the summed signal from all mass channels (*m/z*) along the mass spectral dimension. While these methods incorporate estimates of the background noise and can detect hundreds of analytes, there is growing belief that an undetected fraction of analytes exist in the noise [31]. As a result, signal from one or a few *m/z* can be selected to create an extracted ion current (EIC) chromatogram to facilitate data pre-processing. EICs are useful for well-understood samples, allowing analysts to focus on *m/z* that provide specific information regarding the types of compounds in the sample [32]. Use of an EIC chromatogram can also

allow analysts to find a few peaks that could not be observed in the TIC [33,34]. However, an EIC can potentially lose informative ions that can aid in analyte discovery for insufficiently characterized samples [32]. Therefore, there is a need for an untargeted peak detection algorithm that will not only find these hidden analytes but also enhance their signal against the background noise.

To address this need, we describe the development of a peak detection algorithm that enhances the visualization of signal produced by peaks that have been obscured the background noise using the entire mass spectral dimension collected. A noise threshold is chosen for each *m/z* to ensure that only regions containing chemical signals are identified. By interactively choosing a noise threshold for each *m/z*, our peak detection algorithm accounts for the differences in noise and therefore reduces the identification of false positives. The regions of noise contributions on each *m/z* are zeroed, enhancing the signals of previously undetectable peaks. After each *m/z* undergoes this peak detection algorithm, the signal can be summed together to generate an "enhanced" TIC. Herein, we present the advantages of the enhanced TIC algorithm on a 90-component test mixture at two different concentrations, and on a separation of a yeast cell metabolite extract. This peak detection method is also rigorously evaluated in the context of statistical overlap theory (SOT), referred to as statistical model of overlap. SOT was developed by Davis and Giddings to describe the degree of analyte overlap and the nature of its dependence on peak capacity [35-37]. Ultimately, the SOT provides chromatographers with an understanding between the number of observed peaks and the number of analyte components. Using simulated chromatograms with different noise levels and degrees of saturation, we will show how the enhanced TIC algorithm provides estimates of the total number of peaks that are more consistent with SOT.

8.2. Theory

Originally developed for 1D chromatography, SOT is based on Poisson statistics with the assumption that each component (*i.e.*, analyte) maximum can be represented by a point on a time axis [35,36]. These points are then randomly and independently assigned throughout the separation space. The result is a chromatogram with observable and distinct concentration pulses, referred to as peaks. The resolution between these peaks can be calculated as

$$R_s = \frac{x_0}{4\sigma} \tag{8.1}$$

where x_0 is the minimum distinguishable distance between the two points, which allows for the two peaks to still be identified, and 4σ is the average peak width-at-base, W_b. From this definition of resolution, peak capacity (n_c) can be defined as

$$n_c = \frac{x_{\text{sep}}}{4\sigma R_s} = \frac{x_{\text{sep}}}{4\sigma} \ (R_s = 1) \tag{8.2}$$

where x_{sep} is the separation distance of the chromatogram, which can either be the entire chromatographic run time or a portion of it. The chromatographic saturation (α), relative to peak capacity, is

$$\alpha = \frac{m}{n_c} \tag{8.3}$$

where m is the number of chemical components. It is important to note that the number of observed peaks does not equal m. SOT estimates m based on the number of peaks observed.

Given the statistical limitations of 1D separations predicted [35] and the development of comprehensive 2D chromatography, SOT was extended for these higher order separations [38-46]. Eq. 8.2 can be used to describe the peak capacity of both the ^{1}D and ^{2}D by replacing x_{sep} and σ with their respective values. Thus, the ideal peak capacity of a 2D chromatogram ($n_{c,2D}$) is described as

$$n_{c,2D} = {}^1n_c \times {}^2n_c \tag{8.4}$$

While corrections can be applied to Eq. 8.4 to account for sampling induced peak broadening [44], the analytes simulated in this report were ideally sampled. Due to this ideal sampling, the 2D chromatographic saturation (α_{2D}) can simply be described by Eq. 8.3 with $n_{c,2D}$ substituted in the denominator. The probability of separating two adjacent 2D peaks (P_{2D}) is

$$P_{2D} = e^{-4\alpha_{2D}} \tag{8.5}$$

This probability can be used to estimate the average number of peaks (singlets, doublets, and triplets) in a 2D chromatogram [38]. The total number of peaks (p) can be expressed as [39]

$$p = \sum_{n=1}^{\infty} P_{2D} = m\frac{4\alpha_{2D}e^{-4\alpha_{2D}}}{1-e^{-4\alpha_{2D}}} \tag{8.6}$$

Dividing by $n_{c,2D}$ gives the total number of peaks resolved by the peak capacity as

$$\frac{p}{n_{c,2D}} = \frac{4{\alpha_{2D}}^2 e^{-4\alpha_{2D}}}{1-e^{-4\alpha_{2D}}} \tag{8.7}$$

The resulting equation, however, does not account for the distribution of peak heights or areas in real, complex samples. These equations were derived only by considering analytes to be a single point within the 2D separation. Therefore, for real samples there is a high likelihood of SOT not accounting for analytes that can be obscured by the background noise. While corrections to SOT are being implemented [47,48], herein we utilize these idealized equations to illustrate how the enhance TIC algorithm provides results consistent to SOT, while the standard TIC falls short.

8.3. Methods and Materials

8.3.1. Experimental chromatograms

Separations of the 90-component test mixture (Table F.1) and the yeast cell metabolite extract were both performed on a GC×GC-TOFMS instrument configured with an Agilent 6890N GC (Agilent Technologies, Palo Alto, CA, USA) and LECO Pegasus III TOFMS (LECO Corporation, St. Joseph, MI, USA) using different modulation techniques. A form of flow

modulation termed dynamic pressure gradient modulation (DPGM) [49] was utilized for the separations of the 90-component test mixture. The separation of the yeast cell extract was collected using thermal modulation. Experimental details on the separation of the test mixture [50] and yeast cell extract [5,6] can be found in their respective articles and in Appendix F. Details on importing the experimental data into Matlab 2019b (Mathworks, Inc., Natick, MA, USA), baseline correction (Figure F.1), and data preprocessing can also be found in the Appendix F.

8.3.2. Simulated chromatograms

All chromatographic simulations, with their parameters listed in Table F.2, were performed in Matlab 2019b using the Parallel Computing Toolbox. Each chromatographic simulation contained 40 analytes (m), which had their mass spectra independently and randomly selected. To mimic a realistic metabolomic sample, analyte mass spectra were obtained at unit resolution from the FiehnLib [51]. Table F.3 lists the chemical species used. The ^{1}D separation time ($^1t_{sep}$) varied between 160 and 16 s while the P_M stayed constant at 1 s. Varying the ^{1}D separation time allowed for studying the enhanced TIC peak detection algorithm at 10 different saturation factors (α_{2D} = 0.1 – 1 by intervals of 0.1). An exponential distribution was used to model the peak areas in the simulation using the exprnd() function in Matlab with a mean of 200. The exponential distributions and peak areas were independently and randomly assigned to each chromatogram and analyte, respectively. Modeling the distribution of peak areas in the simulated chromatograms was based upon experimental results. The data collection rate simulated was 100 Hz.

To simulate GC×GC-TOFMS chromatograms, analyte peak profiles were modeled as Gaussian in both separation dimensions. The ^{1}D analyte profile was defined by a peak width-at-

base (^1w_b) of 4 s and its assigned peak area from the exponential distribution. The modulated ^{2}D peaks for a given ^{1}D analyte profile were generated as a series of Gaussian peaks with ^2w_b of 100 ms and the spacing between peak maxima defined by P_M. The peak heights for the modulated ^{2}D peaks were defined by the value apparent for the ^{1}D peak at each modulated peak maximum location. The ^{1}D and ^{2}D retention times were randomly generated for each analyte. After the peak profile of each analyte was generated, the mass spectrum of a randomly selected analyte was applied as an outer product of the unfolded GC×GC chromatogram (vector) and the mass spectrum vector. The chromatograms of individual analytes were then summed together to create a chromatogram containing 40 components and the overall chromatogram was cut into intervals of P_M and stacked to create a GC×GC-TOFMS data cube.

Next, random Gaussian-distributed noise was generated independently for each chromatographic data point on each simulated *m/z*, 40 – 400. The standard deviation of the noise (s_N) was chosen to produce the *S/N* values listed in Table F.2 for a peak with an area of 200 (*i.e.*, relative *S/N*; S/N_{rel}), the mean peak area of the exponential distribution. Figure F.2 shows four different S/N_{rel} (100, 10, 1, and 0.1) for peak with an area of 200, illustrating the baseline noise in simulated chromatograms. Also, to readily apply in terms of the enhanced TIC algorithm, the *S/N* is defined as the ratio of the tallest ^{2}D peak to 6 × s_N. For the simulation study, 100 chromatograms at each α_{2D} (10 total) with four different S/N_{rel} values (100, 10, 1, and 0.1) were generated to yield 4,000 unique chromatograms. Additionally, 100 chromatograms with α_{2D} of 0.1 at the 20 S/N_{rel} values listed in Table F.2 (2,000 unique chromatograms) were also generated.

8.3.3. Enhanced TIC algorithm

Development and application of the enhanced TIC algorithm to both the experimental and simulated chromatograms were performed in Matlab 2019b. First, experimentally collected

chromatograms were baseline corrected using a rolling ball minimum [18,52] and their baselines were re-centered at zero (see Supporting Information). This step was not taken with the simulated chromatograms since lower frequency noise was not included in the simulation. Starting with the chromatograms in their unfolded, vector form, the algorithm interactively calculates the s_N on each *m/z*. An appropriate multiple of s_N for each *m/z* is selected as the signal threshold. The algorithm locates the ^{2}D peak maxima that are larger than the signal threshold set for each *m/z*. To ensure the whole peak shape is preserved, a data window centered around the ^{2}D peak maxima is determined based on the average ^2w_b. For example, the average ^2w_b for the 90-component test mixtures was 130 ms and the data collection rate was 200 Hz (1 data point = 5 ms), so each data window was defined by 26 data points [50]. Similarly, the data windows used for the yeast cell extract and simulated chromatograms were 15 and 10 data points, respectively. All signals per-*m/z* not identified as part of detected ^{2}D peaks are part of the noise and set to zero. Finally, these processed chromatograms can be re-folded to create the GC×GC-TOFMS data cube and summed to generate the enhanced TIC. A schematic of this process is provided in Figure F.3.

8.4. Results and Discussion

8.4.1. Application to experimental chromatograms

Figure 8.1 compares the unfolded GC×GC standard TIC (A), EICs (B), and enhanced TIC (C) of the 10 ppm sample of the 90-component test mixture. To define, the standard TIC is produced by summing the signal at each time point from all *m/z* after baseline correction. Likewise, the enhanced TIC is produced by summing the signal at each time point from all *m/z* after performing the described peak detection algorithm. The inset in Figure 8.1A magnifies a section of background noise for the standard TIC, where its intensity is approximately 2400 due

to the noise component from each *m/z* being added together. Compared to the intensity of the background noise, many of the ^{2}D peaks from the 10 ppm test mixture can be seen in the standard TIC because their summed intensities are larger than the background noise. Figure 8.1B shows an overlay of each EIC and the inset highlights that the intensity of the background noise for a representative EIC is around 30, which is approximately 80-fold lower than the intensity of the background noise in the standard TIC. Therefore, peaks with lower signal that are undiscoverable in the standard TIC can be observed with an EIC. This observation suggests an anticipated benefit to "comprehensively" use all EICs through the enhanced TIC algorithm. To do so, an appropriate signal threshold based on the s_N was determined for each *m/z*. Since previous work [50] indicated there are 58 modulated peaks (out of 90 possible analytes due to several analytes severely overlapping) in the separation highlighted in Figure 8.1A, an appropriate signal threshold for each *m/z* was chosen by ensuring that the maximum number of peaks were filtered through without including signal due to the background noise, *i.e.*, false positives. Figure F.4 illustrates that a threshold of $6 \times s_N$ was able to strike a balance between both variables. The use of this signal threshold is also supported by the fact that the approximate height of random Gaussian noise is 6 to $8 \times s_N$ [53], so signals that are taller than $6 \times s_N$ for a *m/z* are kept because they are due to analyte response and not instrumental noise. Likewise, signals lower than this threshold on each *m/z* are due to instrumental noise and are subsequently zeroed. This step ensures signal enhancement of these lower intensity peaks when the mass spectral dimension is summed together to form the enhanced TIC. Application of the enhanced TIC algorithm using a threshold of $6 \times s_N$ is demonstrated in Figure 8.1C. A higher threshold could be set if the s_N appears to be changing along either the chromatographic or mass spectral

dimensions. However, the s_N observed was essentially constant with respect to separation time (Figure 8.1) and *m/z* collected (Figure F.5).

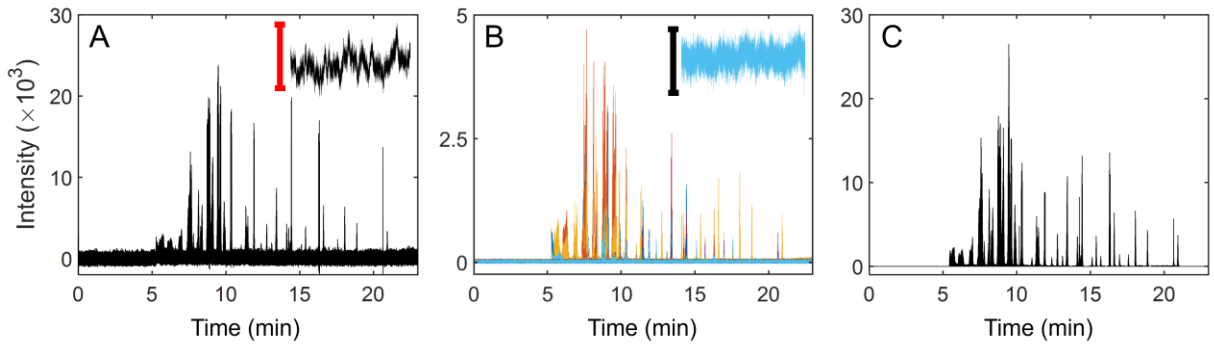


Figure 8.1. Application of the enhanced TIC algorithm to the 10 ppm 90-component test mixture. (A) The unfolded GC×GC data standard TIC for the test mixture after baseline correction. Inset: Zoom-in on the raw baseline noise between 0 – 2 min. The red bar represents an intensity scale of ~ 2400. (B) Chromatogram with the EIC signal traces from all *m/z*. Inset: Zoom-in on the raw baseline noise on an individual *m/z* between 0 – 2 min. The black bar represents an intensity scale of ~ 30. (C) The unfolded GC×GC data enhanced TIC.

Figure 8.2 magnifies a region of the unfolded GC × GC data to compare the standard TIC (i) and enhanced TIC (ii) for the 10 ppm (A; left column) and 1 ppm (B; right column) test mixtures. At 10 ppm in the standard TIC, two prominent analyte ^{2}D peak patterns are observed above the background noise: hexadecane (HX) and eicosane (EC). However, it is difficult to determine if the detected spikes right before 1020 and 1060 s are due to analyte signal. After application of the enhanced TIC algorithm, it is confirmed that the signals at 1020 and 1060 s were also due to analytes. Meanwhile, for the 1 ppm test mixture, none of the analytes known to elute in this time window were observed in the standard TIC. After application of the enhanced TIC algorithm, only HX and EC could be observed. Since 2-dodecanone (DO) and methyl laurate (ML) are observed to have a lower detector response than HX and EC in the 10 ppm test mixture, subsequent dilution to 1 ppm put DO and ML just below the signal threshold of the enhanced TIC. Overall, Figure 8.2 illustrates both the benefit and anticipated limitation of the enhanced TIC algorithm for peak discovery.

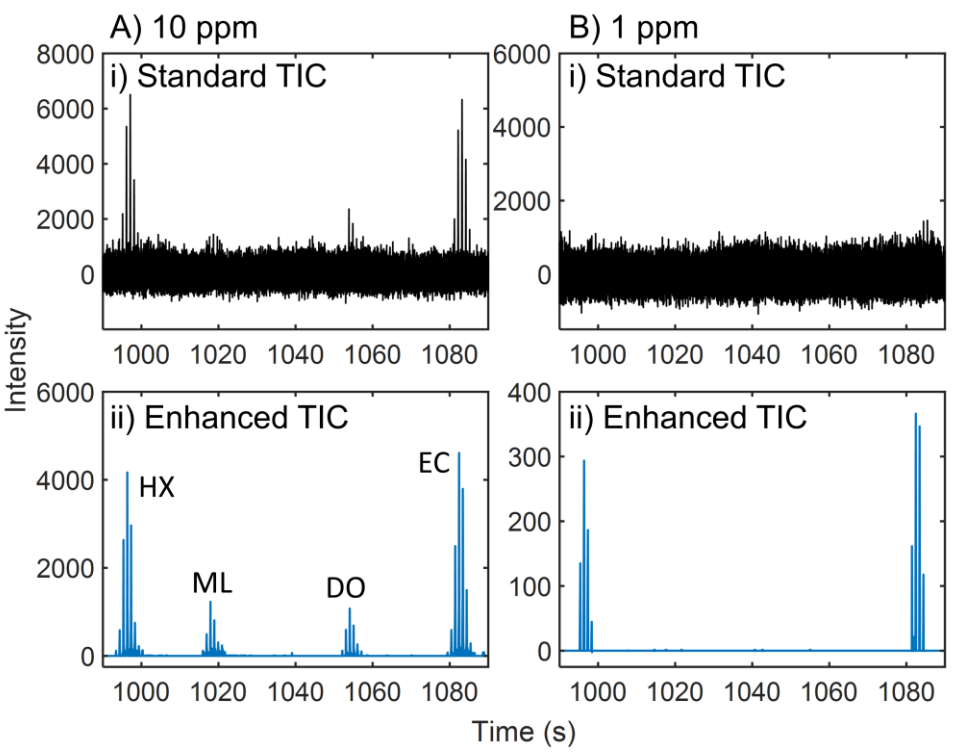


Figure 8.2. Comparison of a zoomed in region from Figure 8.1 of the standard TIC (i) and enhanced TIC (ii) for the 10 ppm (A) and 1 ppm (B) test mixture. Compounds labeled: HX – hexadecane; DO – 2-dodecanone; ML – methyl laurate; EC – eicosane.

For better visualization of the improvement to peak detection, 2D contour plots following enhanced TIC processing for the 10 ppm (A; left column) and 1 ppm (B; right column) ppm test mixtures are provided (Figure 8.3). The chromatograms in the top row are the standard TICs (i) and the bottom row shows the enhanced TICs (ii) with the number of peaks found using either a conventional watershed-based algorithm for the standard TIC or the algorithm described herein for the enhanced TIC are indicated on each chromatogram. Approximately 52 % more peaks were found after using the enhanced TIC algorithm on the 10 ppm sample. Since 58 peaks were observed in the concentrated (10 ppth) test mixture [50], the enhanced TIC recovered 93 % of the original peaks. Meanwhile, the enhanced TIC algorithm was able to find 36 peaks in the 1 ppm test mixture even though none of them could be seen in the standard TIC. The peak recovery at 1 ppm after application of the enhanced TIC was impressively 62 %. Since the

application of the enhanced TIC algorithm was able to highlight previously undiscoverable peaks for these ideal mixtures, the next step was to apply this algorithm to a real, complex sample.

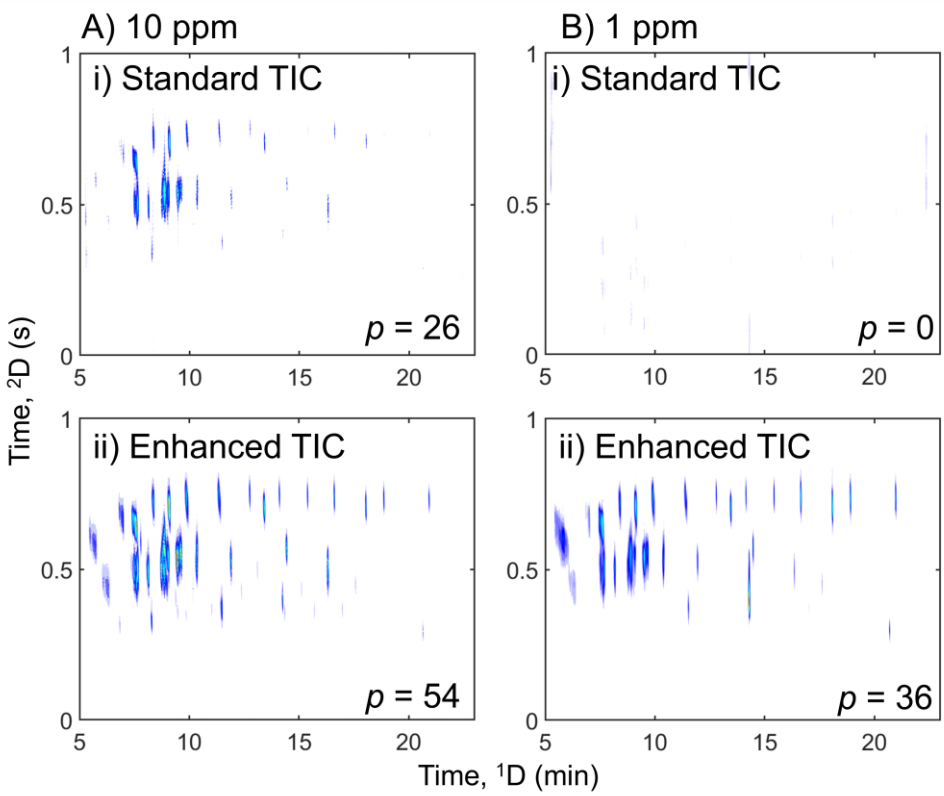


Figure 8.3. Comparison of the standard TICs (i) and enhanced (ii) TICs for the GC×GC separation of the 10 ppm (A) and 1 ppm (B) test mixtures. The unfolded data for 10 ppm is provided in Figure 8.1A for (i) and 1C for (ii). The number of observed peaks (*p*) are stated on each chromatogram.

The enhanced TIC algorithm was also performed on separation of a yeast cell extract (Figure F.6A) [5,6]. Figure 8.4 compares the standard TICs (i) and enhanced TICs (ii) for Sections 1 (A) and 2 (B) of the separation. For Section 1, the enhanced TIC provided a 33 % increase in detected peaks, while with Section 2 there was a 64 % increase. These results illustrate the efficacy of the enhanced TIC algorithm on natural, complex samples. Since the enhanced TIC algorithm fuses the advantages of using both the TIC and EIC, it is useful to compare the enhanced TIC to EICs. Figure F.6 shows the separation of the yeast cell extract as

three EICs: *m/z* 73 (B), *m/z* 205 (C), and *m/z* 387 (D). These *m/z* are selective towards analytes that are trimethylsilyl adducts from the derivatization (B), trimethylsilyl carbohydrates (C), or trimethylsilyl sugar phosphates (D) [5,6]. While plotting one or a few *m/z* in an EIC also allows an analyst to see some of these hidden peaks, the benefit of the enhanced TIC is that it can be applied "comprehensively" to maximize peak discovery without prior knowledge of the sample.

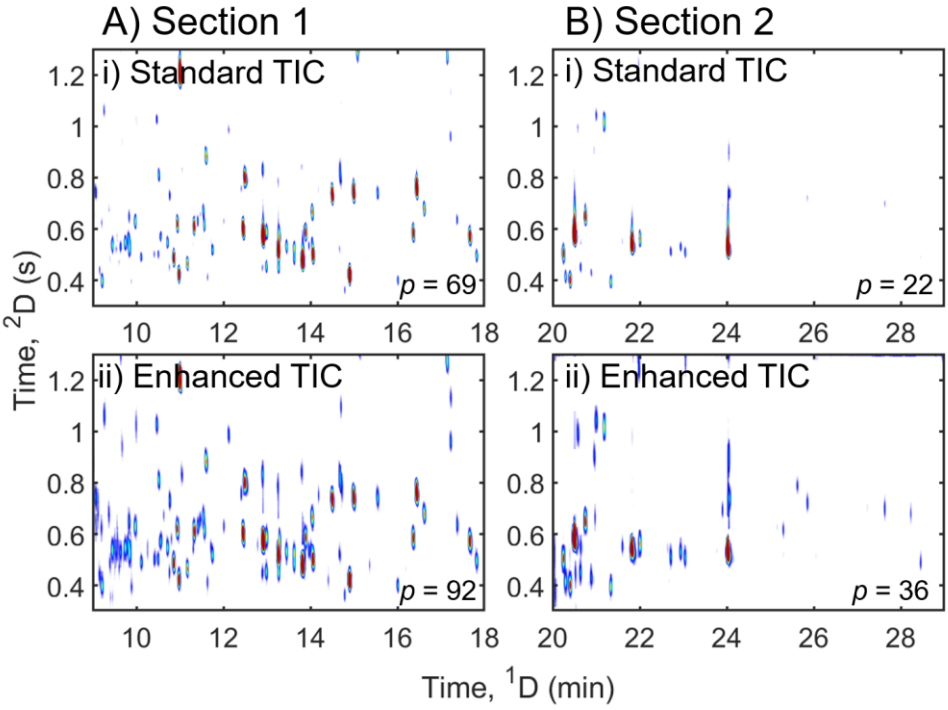


Figure 8.4. Comparison of the standard TICs (i) and enhanced TICs (ii) for Sections 1 (A) and 2 (B) of a separation of a yeast cell extract shown in Figure F.6A. The number of observed peaks (*p*) are stated on each chromatogram.

Figure 8.5 illustrates the number of peaks observed in the enhanced TIC as a function of peak height for the 90-component test mixtures (A) and the yeast metabolome sample (B). The peak height was measured using the tallest ^{2}D peak and the vertical black line indicates the *LOD* for the standard TIC. In Figure 8.5A, for the data at 10 ppm, the number of peaks increase as the peak height threshold decreases from the *LOD* of the standard TIC (~ 2400; Figure 8.1A inset) to the 6 × s_N threshold used to describe the noise intensity on each *m/z* (~ 30; Figure 8.1B inset). As

for the 1 ppm sample, note that the respective peak heights are all below the *LOD* for the standard TIC, but above the intensity of the background noise on each EIC. The inset in Figure 8.5 emphasizes this result for the peak height distribution at 1 ppm. With peak intensities ranging from ~ 30 to 500, all peaks at 1 ppm were originally obscured in the standard TIC by the background noise intensity. A similar trend is seen in Figure 8.5B for Sections 1 and 2 of the yeast cell extract separation, where the number of peaks observed increases from the *LOD* in the standard TIC to the signal threshold determined by the enhanced TIC algorithm for each *m/z*.

The true distribution of peak heights and/or concentrations in complex samples has been unknown because the number of peaks with a signal below the *LOD* was undeterminable [31]. Use of the enhanced TIC algorithm could allow analysts to gain more information regarding these distributions. Previous research has described these distributions as either an exponential [54-46] or log-normal [31,48,57] function based on their analytical response. Here, an exponential distribution (dashed lines) was found to best describe the peak height distributions shown in Figure 8.5. Both the coefficient of determination (r^2) and lack of fit (*LOF*) [50,58] were used to describe the goodness-of-fit for this model. The exponential fit for the 10 ppm sample had a r^2 of 0.98 and *LOF* of 9.1 %. Similarly, the exponential fit for the 1 ppm sample had a r^2 of 0.99 and *LOF* of 6.9 %. A more uniform peak height distribution for the test mixtures would be expected since all the analytes are at the same concentration. However, variation in analyte fragmentation, analyte interactions, and differences in modulation sampling coupled with the enhanced TIC algorithm using a signal threshold can cause the non-uniform distribution seen in Figure 8.5. For example, an analyte with a more complex mass spectrum could have more signal distributed below the signal threshold set by the enhanced TIC than one with a simpler mass spectrum. To better understand the signal response distribution of a natural sample, exponential

functions were also fitted to the yeast cell extract data (Figure 8.5B). For Section 1, the exponential fit had a r^2 of 0.98 and *LOF* of 8.8 %. The exponential fit for Section 2 had a r^2 of 0.98 and *LOF* of 8.5 %. These results illustrate how the enhanced TIC algorithm may provide additional insight into the study of response distributions of probe and natural mixtures.

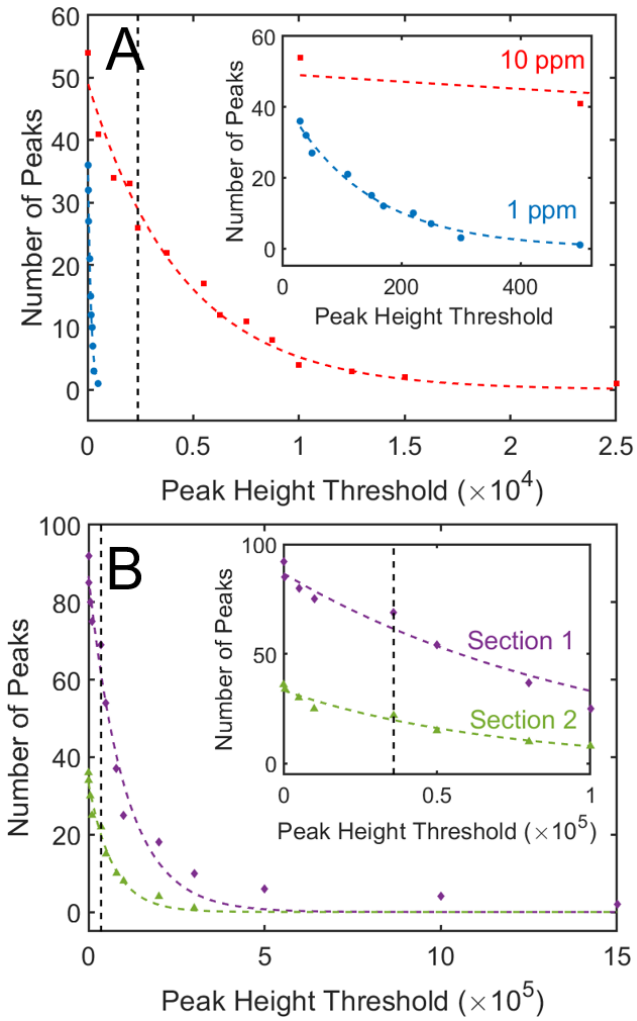


Figure 8.5. The number of peaks in the enhanced TIC as a function of the peak height threshold for both concentrations (10 ppm – red squares; 1 ppm – blue circles) of the test mixture (A) and sections (Section 1 – purple diamonds; Section 2 – green triangles) of the yeast cell extract data (B). The vertical dashed black line indicates the height of the noise in the standard TIC. The dashed lines are fitted to exponential functions. Inset: Zoom-in of the lower peak height thresholds for emphasis.

8.4.2. Application to simulated chromatograms

To further understand the benefits of using the enhanced TIC algorithm, chromatographic simulations based on the principles of SOT were executed. Figure 8.6 compares the standard TIC (top row; i) and enhanced TIC (bottom row; ii) of a representative simulation of 40 analytes (α_{2D} = 0.1) with a S/N_{rel} of 100 (A), 10 (B), 1 (C), and 0.1 (D). The exponential peak area distribution for these representative chromatograms is shown in Figure F.7. Since the exponential distribution for these simulations used a mean peak area of 200, this was translated into the *S/N* for each chromatogram (Figure F.2). SOT predicts that the maximum number of peaks observed in a chromatogram at α_{2D} of 0.1 would be ~ 33 peaks. At a S/N_{rel} of 100, both the enhanced and standard TICs are close to this maximum peak value, so there is not a significant advantage in using the enhanced TIC algorithm. However, the enhanced TIC did detect 26 % more peak than the standard TIC at a S/N_{rel} of 10 (Figure 8.6B). Use of the enhanced TIC algorithm also detected approximately 4 and 5 times as many peaks for S/N_{rel} of 1 and 0.1, respectively (Figure 8.6C-D). However, Figure 8.6D illustrates that the enhanced TIC algorithm is also ultimately limited by the background noise. If the background noise is significantly larger than the average of the peak height distribution of the chromatogram, than the enhanced TIC algorithm will only enhance analyte signals whose concentration were skewed to the right of the average peak height (Figure F.7). Therefore, the dependence of the algorithm on the background noise needs to be considered when interpreting these results.

The results presented in Figure 8.6 only highlight one representative chromatogram at one α_{2D}, so more analyses were undertaken for 100 independently simulated chromatograms at the four different S/N_{rel} values with α_{2D} ranging from 0.1 to 1. The average number of peaks detected in both the standard TICs and enhanced TICs as a function of α_{2D} at different S/N_{rel}

levels are provided in Figure F.8. Briefly, the enhanced TIC algorithm identifies more peaks than the standard TIC at all S/N_{rel} levels, except for S/N_{rel} of 100. At a S/N_{rel} of 100, essentially all the analytes are readily detectable by either the standard TIC or enhanced TIC. The benefit of using the enhanced TIC is more pronounced at lower S/N_{rel} levels. For example, at a S/N_{rel} of 10, the enhanced TIC detects an average of 17 – 47 % more peaks than the standard TIC for all α_{2D} investigated. For a S/N_{rel} of 1 and 0.1, the enhanced TIC increased peak detection by a factor of ~ 3 and ~ 4, respectively. It should also be noted that the number of peaks detected with both the standard TICs and enhanced TICs decreases as α_{2D} increases, since more analytes overlap with one another, resulting in fewer peaks per chromatogram. Accordingly, there was a need to compare the results presented in Figure F.8 in the context of the SOT.

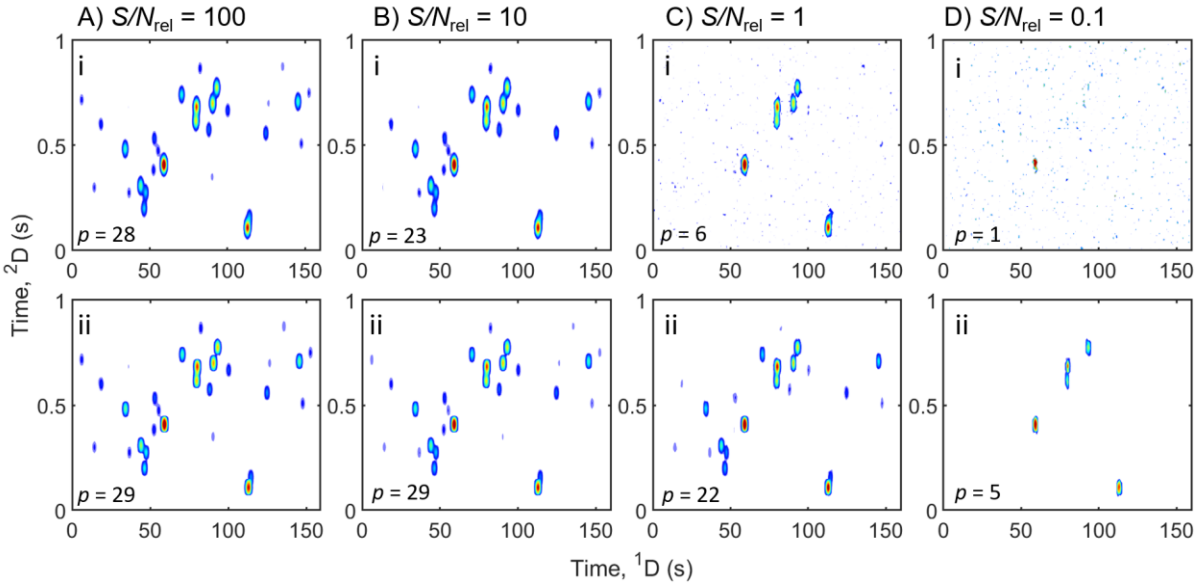


Figure 8.6. Comparison of standard TIC (i) and enhanced TIC (ii) for a representative simulated chromatogram at four different relative signal-to-noise (S/N_{rel}) values: (A) 100, (B) 10, (C) 1, (D) 0.1. The saturation factor (α) simulated was 0.1 (40 components in a peak capacity space of 400). The number of peaks (p) are stated on each chromatogram. SOT predicts that the maximum number of peaks observed would be 33.

SOT allows for estimations of the total number of peaks in a separation based on the probability of resolving two adjacent chromatographic peaks (Eq. 8.6) [39]. For this purpose, the

results for peaks, p, in Figure F.5 were normalized to the 2D peak capacity, $n_{c,2D}$, of the separation for comparisons between different α_{2D} (Eq. 8.7). Thus, Figure 8.7 illustrates the average number of peaks per peak capacity ($p/n_{c,2D}$) for the standard TIC (blue circles) and enhanced TIC (red squares) as a function of α_{2D} for the different simulated S/N_{rel} values: 100 (A), 10 (B), 1 (C), and 0.1 (D). The black line on the plot represents the relationship predicted by SOT (Eq. 8.7). Furthermore, the *LOF* was calculated to quantify the degree of deviation between SOT and the standard TICs and enhanced TICs (Table F.4). Overall, visual inspection shows that the standard TIC begins to deviate at a higher S/N_{rel} than the enhanced TIC. First, the $p/n_{c,2D}$ for both the standard and enhanced TICs are true to the predictions made with SOT at a S/N_{rel} of 100 with respective *LOF*s of 5.0 and 2.5 %. At a S/N_{rel} of 10, the standard TIC begins to deviate from the predictions made by SOT with a *LOF* of 25.2 %. However, the *LOF* between SOT and the results of the enhanced TIC at a S/N_{rel} of 10 is 5.2 %. These results emphasize the advantage of using the enhanced TIC for chromatograms with low *S/N* levels. Beginning at a S/N_{rel} of 1, the enhanced TIC begins to deviate from the predictions made by SOT. The *LOF* for the standard TIC and enhanced TIC was 68.0 and 20.0 %, respectively, at a S/N_{rel} of 1. Lastly, for S/N_{rel} of 0.1, a *LOF* of 83.9 and 53.4 % illustrate the large discrepancies between SOT and the standard TIC and enhanced TIC, respectively. Ultimately, these results illustrate that the standard TIC deviates from SOT between a S/N_{rel} of 100 and 10. More advantageously, the enhanced TIC deviates from SOT between a S/N_{rel} of 10 and 1.

Finally, Eq. 8.7 was used to predict the α_{2D} of chromatograms with S/N_{rel} ranging from 100 to 0.1 (Table F.2) before and after application of the enhanced TIC algorithm (Figure F.9). For the standard TIC, the predicted α_{2D} from SOT equals the true α_{2D} of 0.1 until a S/N_{rel} of 50. Similarly, for the enhanced TIC, the predicted α_{2D} equals the true α_{2D} of the simulation until a

S/N_{rel} of 3. These collective results illustrate that the enhanced TIC algorithm provides not only signal enhancement of peaks obscured by the background noise, but also results consistent with SOT for chromatograms with lower S/N. Hence, it would be a valuable tool for studies involving simulated chromatograms.

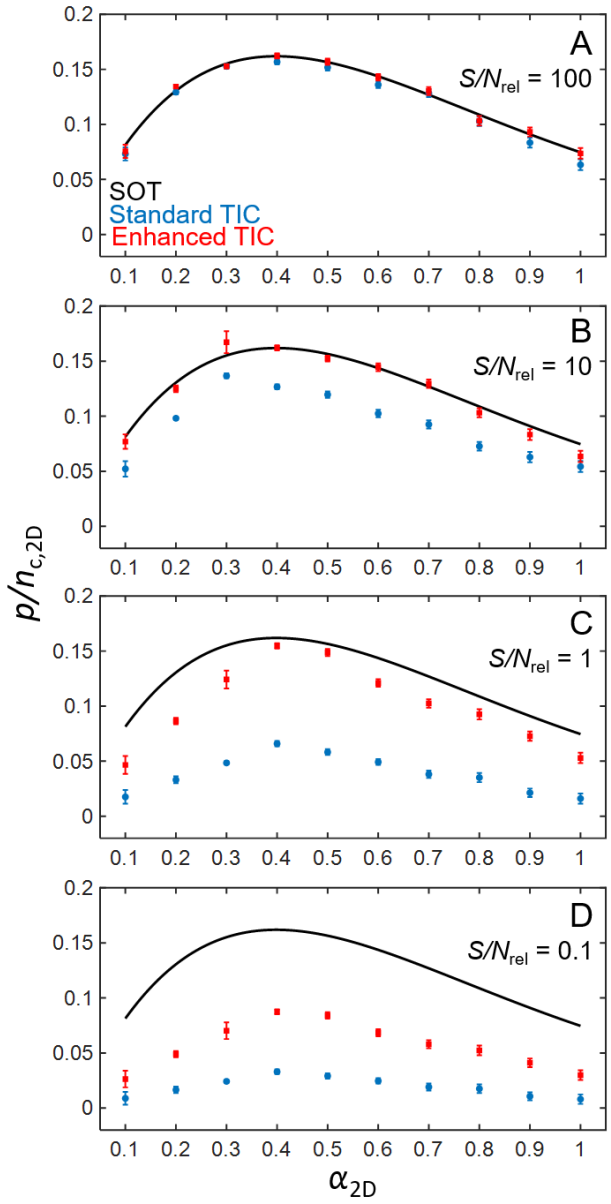


Figure 8.7. The number of peaks per peak capacity ($p/n_{c,2D}$) as a function of α_{2D} simulated at four different S/N_{rel} values: (A) 100, (B) 10, (C) 1, (D) 0.1. Plots show the relationship predicted by SOT applying Eq. 8.7 (black line) and the results for the standard TIC (blue circles) and enhanced TIC (red squares). Results for each point are shown as the average and standard deviations of 100 simulations.

8.5. Conclusion

An untargeted algorithm to find analytes that were previously obscured by the background noise, referred to as the enhanced TIC method, was presented for GC×GC-TOFMS data. This peak detection method was found to identify analyte signals in the background noise with recovery rates of 62 and 93 % for the 1 and 10 ppm test mixtures, respectively. Application of the enhanced TIC algorithm on a yeast cell extract separation was also shown to provide more comprehensive information than use of EICs. It was demonstrated that the enhanced TIC algorithm allows for a deeper understanding of concentration distributions in complex samples and accurate predictions with SOT. While this method was developed for GC×GC, utility appears to be broad in scope, such as 1D GC-MS separations. Future studies will investigate the application of the enhanced TIC algorithm to both 1D and comprehensive 2D liquid chromatography-MS (LC-MS and LC×LC-MS), which presents the challenge of a complex background due to the high chemical noise from the mobile phase additives [59] and ion suppression altering the concentration of analyte ions [60]. Fortunately, the enhanced TIC algorithm allows for modifications to both the signal threshold set on each *m/z* and the preserved data window of each peak in order to de-noise the chromatogram and capture distorted peak shapes. Overall, this unsupervised "comprehensive" peak detection algorithm may serve as an important step in pre-processing prior to chemometric analyses.

8.6. References

[1] Z. Liu, J.B. Phillips, Comprehensive two-dimensional gas chromatography using an on-column thermal modulator interface, J. Chromatogr. Sci. 29 (1991) 227–231. https://doi.org/10.1093/chromsci/29.6.227.

[2] H.D. Bahaghighat, C.E. Freye, R.E. Synovec, Recent advances in modulator technology for comprehensive two dimensional gas chromatography, TrAC - Trends Anal. Chem. 113 (2019) 379–391. https://doi.org/10.1016/j.trac.2018.04.016.

[3] M.S. Klee, J. Cochran, M. Merrick, L.M. Blumberg, Evaluation of conditions of comprehensive two-dimensional gas chromatography that yield a near-theoretical

maximum in peak capacity gain, J. Chromatogr. A. 1383 (2015) 151–159. https://doi.org/10.1016/j.chroma.2015.01.031.

[4] Z. Liu, D.G. Patterson, M.L. Lee, Geometric approach to factor analysis for the estimation of orthogonality and practical peak capacity in comprehensive two-dimensional separations, Anal. Chem. 67 (1995) 3840–3845. https://doi.org/10.1021/ac00117a004.

[5] R.E. Mohler, K.M. Dombek, J.C. Hoggard, E.T. Young, R.E. Synovec, Comprehensive two-dimensional gas chromatography time-of-flight mass spectrometry analysis of metabolites in fermenting and respiring yeast cells, Anal. Chem. 78 (2006) 2700–2709. https://doi.org/10.1021/ac052106o.

[6] R.E. Mohler, K.M. Dombek, J.C. Hoggard, K.M. Pierce, E.T. Young, R.E. Synovec, Comprehensive analysis of yeast metabolite GC×GC-TOFMS data: Combining discovery-mode and deconvolution chemometric software, Analyst. 132 (2007) 756–767. https://doi.org/10.1039/b700061h.

[7] A.C. Beckstrom, E.M. Humston, L.R. Snyder, R.E. Synovec, S.E. Juul, Application of comprehensive two-dimensional gas chromatography with time-of-flight mass spectrometry method to identify potential biomarkers of perinatal asphyxia in a non-human primate model, J. Chromatogr. A. 1218 (2011) 1899–1906. https://doi.org/10.1016/j.chroma.2011.01.086.

[8] M.F. Almstetter, P.J. Oefner, K. Dettmer, Comprehensive two-dimensional gas chromatography in metabolomics, Anal. Bioanal. Chem. 402 (2012) 1993–2013. https://doi.org/10.1007/s00216-011-5630-y.

[9] O. Pani, T. Górecki, Comprehensive two-dimensional gas chromatography (GC×GC) in environmental analysis and monitoring, Anal. Bioanal. Chem. 386 (2006) 1013–1023. https://doi.org/10.1007/s00216-006-0568-1.

[10] E. Skoczyńska, P. Korytár, J. De Boer, Maximizing chromatographic information from environmental extracts by GCxGC-ToF-MS, Environ. Sci. Technol. 42 (2008) 6611–6618. https://doi.org/10.1021/es703229t.

[11] R.E. Mohler, K.T. O'Reilly, D.A. Zemo, A.K. Tiwary, R.I. Magaw, K.A. Synowiec, Non-targeted analysis of petroleum metabolites in groundwater using GC×GC-TOFMS, Environ. Sci. Technol. 47 (2013) 10471–10476. https://doi.org/10.1021/es401706m.

[12] E.M. Humston, J.D. Knowles, A. McShea, R.E. Synovec, Quantitative assessment of moisture damage for cacao bean quality using two-dimensional gas chromatography combined with time-of-flight mass spectrometry and chemometrics, J. Chromatogr. A. 1217 (2010) 1963–1970. https://doi.org/10.1016/j.chroma.2010.01.069.

[13] C. Cordero, J. Kiefl, P. Schieberle, S.E. Reichenbach, C. Bicchi, Comprehensive two-dimensional gas chromatography and food sensory properties: Potential and challenges, Anal. Bioanal. Chem. 407 (2015) 169–191. https://doi.org/10.1007/s00216-014-8248-z.

[14] C. Cordero, J. Kiefl, S.E. Reichenbach, C. Bicchi, Characterization of odorant patterns by comprehensive two-dimensional gas chromatography: A challenge in omic studies, TrAC - Trends Anal. Chem. 113 (2019) 364–378. https://doi.org/10.1016/j.trac.2018.06.005.

[15] G.S. Frysinger, R.B. Gaines, Forensic analysis of ignitable liquids in fire debris by comprehensive two-dimensional gas chromatography, J. Forensic Sci. 47 (2002) 471–482. https://doi.org/10.1520/JFS15288J.

[16] S.M. Song, P. Marriott, A. Kotsos, O.H. Drummer, P. Wynne, Comprehensive two-dimensional gas chromatography with time-of-flight mass spectrometry (GC × GC-TOFMS) for drug screening and confirmation, Forensic Sci. Int. 143 (2004) 87–101. https://doi.org/10.1016/j.forsciint.2004.02.042.

[17] J.C. Hoggard, J.H. Wahl, R.E. Synovec, G.M. Mong, C.G. Fraga, Impurity Profiling of a Chemical Weapon Precursor for Possible Forensic Signatures by Comprehensive Two-Dimensional Gas Chromatography/Mass Spectrometry and Chemometrics, Anal. Chem. 82 (2010) 689–698. https://doi.org/10.1021/ac902247x.

[18] B. Kehimkar, J.C. Hoggard, L.C. Marney, M.C. Billingsley, C.G. Fraga, T.J. Bruno, R.E. Synovec, Correlation of rocket propulsion fuel properties with chemical composition using comprehensive two-dimensional gas chromatography with time-of-flight mass spectrometry followed by partial least squares regression analysis, J. Chromatogr. A. 1327 (2014) 132–140. https://doi.org/10.1016/j.chroma.2013.12.060.

[19] B.J. Pollo, G.L. Alexandrino, F. Augusto, L.W. Hantao, The impact of comprehensive two-dimensional gas chromatography on oil & gas analysis: Recent advances and applications in petroleum industry, TrAC - Trends Anal. Chem. 105 (2018) 202–217. https://doi.org/10.1016/j.trac.2018.05.007.

[20] S.E. Prebihalo, K.L. Berrier, C.E. Freye, H.D. Bahaghighat, N.R. Moore, D.K. Pinkerton, R.E. Synovec, Multidimensional Gas Chromatography: Advances in Instrumentation, Chemometrics, and Applications, Anal. Chem. 90 (2018) 505–532. https://doi.org/10.1021/acs.analchem.7b04226.

[21] D.W. Cook, S.C. Rutan, Chemometrics for the analysis of chromatographic data in metabolomics investigations, J. Chemom. 28 (2014) 681–687. https://doi.org/10.1002/cem.2624.

[22] M. Navarro-Reig, C. Bedia, R. Tauler, J. Jaumot, Chemometric strategies for peak detection and profiling from multidimensional chromatography, Proteomics. 18 (2018) 1–12. https://doi.org/10.1002/pmic.201700327.

[23] S. Peters, G. Vivó-Truyols, P.J. Marriott, P.J. Schoenmakers, Development of an algorithm for peak detection in comprehensive two-dimensional chromatography, J. Chromatogr. A. 1156 (2007) 14–24. https://doi.org/10.1016/j.chroma.2006.10.066.

[24] A. Bieniek, A. Moga, An efficient watershed algorithm based on connected components, Pattern Recognit. 33 (2000) 907–916. https://doi.org/10.1016/S0031-3203(99)00154-5.

[25] S.E. Reichenbach, M. Ni, V. Kottapalli, A. Visvanathan, Information technologies for comprehensive two-dimensional gas chromatography, Chemom. Intell. Lab. Syst. 71 (2004) 107–120. https://doi.org/10.1016/j.chemolab.2003.12.009.

[26] S.E. Reichenbach, X. Tian, Q. Tao, D.R. Stoll, P.W. Carr, Comprehensive feature analysis for sample classification with comprehensive two-dimensional LC, J. Sep. Sci. 33 (2010) 1365–1374. https://doi.org/10.1002/jssc.200900859.

[27] G. Vivó-Truyols, Bayesian approach for peak detection in two-dimensional chromatography, Anal. Chem. 84 (2012) 2622–2630. https://doi.org/10.1021/ac202124t.

[28] A. Barcaru, G. Vivó-Truyols, Use of Bayesian Statistics for Pairwise Comparison of Megavariate Data Sets: Extracting Meaningful Differences between GCxGC-MS Chromatograms Using Jensen–Shannon Divergence, Anal. Chem. 88 (2016) 2096–2104. https://doi.org/10.1021/acs.analchem.5b03506.

[29] S. Kim, M. Ouyang, J. Jeong, C. Shen, X. Zhang, A new method of peak detection for analysis of comprehensive two-dimensional gas chromatography mass spectrometry data, Ann. Appl. Stat. 8 (2014) 1209–1231. https://doi.org/10.1214/14-AOAS731.

[30] S. Kim, H. Jang, I. Koo, J. Lee, X. Zhang, Normal–Gamma–Bernoulli peak detection for analysis of comprehensive two-dimensional gas chromatography mass spectrometry data, Comput. Stat. Data Anal. 105 (2017) 96–111. https://doi.org/10.1016/j.csda.2016.07.015.

[31] C.G. Enke, L.J. Nagels, Undetected components in natural mixtures: How many? What concentrations? Do they account for chemical noise? What is needed to detect them?, Anal. Chem. 83 (2011) 2539–2546. https://doi.org/10.1021/ac102818a.

[32] L.A. Adutwum, J.J. Harynuk, Unique Ion Filter: A Data Reduction Tool for GC/MS Data Preprocessing Prior to Chemometric Analysis, Anal. Chem. 86 (2014) 7726–7733. https://doi.org/10.1021/ac501660a.

[33] W. Zhang, P.X. Zhao, Quality evaluation of extracted ion chromatograms and chromatographic peaks in liquid chromatography/mass spectrometry-based metabolomics data, BMC Bioinformatics. 15 (2014) 1–13. https://doi.org/10.1186/1471-2105-15-S11-S5.

[34] O.D. Myers, S.J. Sumner, S. Li, S. Barnes, X. Du, One Step Forward for Reducing False Positive and False Negative Compound Identifications from Mass Spectrometry Metabolomics Data: New Algorithms for Constructing Extracted Ion Chromatograms and Detecting Chromatographic Peaks, Anal. Chem. 89 (2017) 8696–8703. https://doi.org/10.1021/acs.analchem.7b00947.

[35] J.M. Davis, J.C. Giddings, Statistical theory of component overlap in multicomponent chromatograms, Anal. Chem. 55 (1983) 418–424. https://doi.org/10.1021/ac00254a003.

[36] J.M. Davis, J.C. Giddings, Origin and characterization of deparatures from the statistical model of component-peak overlap in chromatography, J. Chromatogr. 289 (1984) 277–298.

[37] J.M. Davis, J.C. Giddings, Statistical method for estimation of number of components from single complex chromatograms: Theory, computer-based testing, and analysis of errors, Anal. Chem. 57 (1985) 2168–2177. https://doi.org/10.1021/ac00289a002.

[38] J.M. Davis, Statistical theory of spot overlap in two-dimensional separations, Anal. Chem. 63 (1991) 2141–2152. https://doi.org/10.1021/ac00019a014.

[39] F.J. Oros, J.M. Davis, Comparison of statistical theories of spot overlap in two-dimensional separations and verification of means for estimating the number of zones, J. Chromatogr. 591 (1992) 1–18.

[40] W. Shi, J.M. Davis, Test of theory of overlap for two-dimensional separations by computer simulations of three-dimensional concentration profiles, Anal. Chem. 65 (1993) 482–492. https://doi.org/10.1021/ac00052a028.

[41] J.M. Davis, Statistical theory of spot overlap for n-dimensional separations, Anal. Chem. 65 (1993) 2014–2023. https://doi.org/10.1021/ac00063a015.

[42] J.M. Davis, Statistical-overlap theory for elliptical zones of high aspect ratio in comprehensive two-dimensional separations, J. Sep. Sci. 28 (2005) 347–359. https://doi.org/10.1002/jssc.200401798.

[43] S. Liu, J.M. Davis, Dependence on saturation of average minimum resolution in two-dimensional statistical-overlap theory: Peak overlap in saturated two-dimensional separations, J. Chromatogr. A. 1126 (2006) 244–256. https://doi.org/10.1016/j.chroma.2006.05.064.

[44] J.M. Davis, D.R. Stoll, P.W. Carr, Effect of first-dimension undersampling on effective peak capacity in comprehensive two-dimensional separations, Anal. Chem. 80 (2008) 461–473. https://doi.org/10.1021/ac071504j.

[45] J.M. Davis, D.R. Stoll, P.W. Carr, Dependence of effective peak capacity in comprehensive two-dimensional separations on the distribution of peak capacity between the two dimensions, Anal. Chem. 80 (2008) 8122–8134. https://doi.org/10.1021/ac800933z.

[46] J.M. Davis, P.W. Carr, Effective saturation: A more informative metric for comparing peak separation in one- and two-dimensional separations, Anal. Chem. 81 (2009) 1198–1207. https://doi.org/10.1021/ac801728k.

[47] F. Dondi, A. Bassi, A. Cavazzini, M.C. Pietrogrande, A Quantitative Theory of the Statistical Degree of Peak Overlapping in Chromatography, Anal. Chem. 70 (1998) 766–773. https://doi.org/10.1021/ac9705430.

[48] J.M. Davis, New theory for distribution of minimum resolution in multi-component separations with noise/detection limits, J. Chromatogr. A. 1251 (2012) 1–9. https://doi.org/10.1016/j.chroma.2012.06.034.

[49] T.J. Trinklein, D. V. Gough, C.G. Warren, G.S. Ochoa, R.E. Synovec, Dynamic pressure gradient modulation for comprehensive two-dimensional gas chromatography, J. Chromatogr. A. 1609 (2020). https://doi.org/10.1016/j.chroma.2019.460488.

[50] S. Schöneich, D. V. Gough, T.J. Trinklein, R.E. Synovec, Dynamic pressure gradient modulation for comprehensive two-dimensional gas chromatography with time-of-flight mass spectrometry detection, J. Chromatogr. A. 1620 (2020) 460982. https://doi.org/10.1016/j.chroma.2020.460982.

[51] T. Kind, G. Wohlgemuth, D.Y. Lee, Y. Lu, M. Palazoglu, S. Shahbaz, O. Fiehn, FiehnLib: Mass spectral and retention index libraries for metabolomics based on quadrupole and time-of-flight gas chromatography/mass spectrometry, Anal. Chem. 81 (2009) 10038–10048. https://doi.org/10.1021/ac9019522.

[52] H.G. Schmarr, J. Bernhardt, Profiling analysis of volatile compounds from fruits using comprehensive two-dimensional gas chromatography and image processing techniques, J. Chromatogr. A. 1217 (2010) 565–574. https://doi.org/10.1016/j.chroma.2009.11.063.

[53] S.W. Smith, Statistics, Probability and Noise, in: Sci. Eng. Guid. to Digit. Signal Process. Stat. Probab. Noise, 2nd ed., California Technical Publishing, 1999: pp. 11–34.

[54] D.P. Herman, M.F. Gonnord, G. Guiochon, Statistical approach for estimating the total number of components in complex mixtures from nontotally resolved chromatograms, Anal. Chem. 56 (1984) 995–1003. https://doi.org/10.1021/ac00270a030.

[55] F. Dondi, Y. Du Ale Kahie, G. Lodi, M. Remelli, P. Reschiglian, C. Bighi, Evaluation of the number of components in multi-component liquid chromatograms of plant extracts, Anal. Chim. Acta. 191 (1986) 261–273. https://doi.org/10.1016/S0003-2670(00)86313-8.

[56] M.R. Schure, The dimensionality of chromatographic separations, J. Chromatogr. A. 1218 (2011) 293–302. https://doi.org/10.1016/j.chroma.2010.11.016.

[57] A. Gundlach-Graham, C.G. Enke, Effect of response factor variations on the response distribution of complex mixtures, Eur. J. Mass Spectrom. 21 (2015) 471–479. https://doi.org/10.1255/ejms.1369.

[58] Y. Izadmanesh, E. Garreta-Lara, J.B. Ghasemi, S. Lacorte, V. Matamoros, R. Tauler, Chemometric analysis of comprehensive two dimensional gas chromatography–mass spectrometry metabolomics data, J. Chromatogr. A. 1488 (2017) 113–125. https://doi.org/10.1016/j.chroma.2017.01.052.

[59] S. Chatterjee, G.H. Major, B. Paull, E.S. Rodriguez, M. Kaykhaii, M.R. Linford, Using pattern recognition entropy to select mass chromatograms to prepare total ion current chromatograms from raw liquid chromatography–mass spectrometry data, J. Chromatogr. A. 1558 (2018) 21–28. https://doi.org/10.1016/j.chroma.2018.04.042.

[60] F. Gosetti, E. Mazzucco, D. Zampieri, M.C. Gennaro, Signal suppression/enhancement in high-performance liquid chromatography tandem mass spectrometry, J. Chromatogr. A. 1217 (2010) 3929–3937. https://doi.org/10.1016/j.chroma.2009.11.060.

Chapter 9: Enhancing Gas Chromatography-Mass Spectrometry Resolution and Pure Analyte Discovery using Intra-Chromatogram Elution Profile Matching

9.1. Introduction

Gas chromatography-mass spectrometry (GC-MS) is a fundamental analytical technique for the separation of volatiles and semi-volatiles in various application areas, including petroleomics [1–4], metabolomics [5–8], and food analysis [9–12]. However, as research goals have shifted from targeted to non-targeted studies, the separation of all analytes in a complex matrix with GC-MS has been increasingly challenged due to the limited peak capacity of a one-dimensional GC-MS separation. This idea was demonstrated by Davis and Giddings in 1983 with their work on statistical overlap theory (SOT) [13]. Predictions made with SOT demonstrated that the maximum number of resolvable peaks (i.e., a measurable concentration pulse due to either a single analyte or multiple analytes) in a chromatogram equaled 37 % of the peak capacity [13]. More critically, the maximum number of resolvable, single-analyte peaks was limited to 18 % of the peak capacity [13]. The inevitable peak overlap that occurs in a GC-MS separation creates a difficult situation for analyte identification and quantitation.

For these chromatographic regions with excessive peak overlap, the use of chemometrics is beneficial in mathematically resolving overlapped analytes to improve identification and quantitation efforts. Various chemometric decomposition methods exist to model the pure instrumental response from different analytes using the originally collected data [14]. For GC-MS data, the instrumental response measured for an analyte is a matrix, obtained by computing the outer product of its chromatographic elution profile and mass spectrum. This second-order data structure enables the use of bilinear chemometric decomposition methods like multivariate

curve resolution-alternating least squares (MCR-ALS) [15]. Application of an MCR-ALS model on a region in a GC-MS chromatogram should ideally provide the pure elution profiles and resolved mass spectrum for each analyte. Using these outputs, analyte identification can be performed by matching the retention time (t_R) measured in the pure elution profiles and/or the resolved mass spectrum to a library developed with analytical standards. Additionally, quantitative information about the relative concentration of each analyte can also be determined from the purified elution profiles provided by the MCR-ALS model.

However, the presence of model ambiguities and significant signal contributions from noise/interfering signals can affect the accuracy of an MCR-ALS model. Model ambiguities arise when multiple solutions to the model are feasible (i.e., equally fit the data) [16,17]. The most problematic ambiguity in an MCR-ALS is rotational, which affects the decomposed elution profiles and mass spectra [16,18]. Hence, the presence of rotational ambiguities can cause large errors in analyte identification and quantitation. Along with this modeling ambiguity, signal contributions from noise and overlapped species (i.e., interferences) can also contribute to the poor performance of MCR-ALS. Previous work has demonstrated that the accuracy of MCR-ALS decreases as chromatographic resolution (R_s) decreases, spectrum contamination from noise and larger interferent signals increases, and the similarity between overlapped analytes increases [19–23]. For instance, a minimum R_s of ~ 0.3 was required to achieve confident analyte identification and adequate quantitation error with MCR-ALS [21]. We assert that the unsatisfactory results provided by MCR-ALS under these conditions is also due to the "chemometric multiplex disadvantage" [22,23]. This phenomenon is conceptually akin to the multiplex disadvantage observed for spectroscopic instruments with simultaneous collection, where large signals at one wavelength hamper the measurement of small signals at another

wavelength [24]. In a similar fashion, decomposition models like MCR-ALS analyzes the entire spectra domain simultaneously, and in an effort to minimize the residual error, signals from noise and/or interfering compounds to be included into the results for the target species [22,23].

Fortunately, both rotational ambiguities and the chemometric multiplex disadvantage can be overcome by providing MCR-ALS with the pure elution profiles and/or mass spectra for the analytes in the model [19,22,23,25–28]. Estimates of these pure elution profiles and/or mass spectra can be determined either experimentally, with analytical standards [29,30], or mathematically, with multivariate computational procedures [31,32]. Herein, we present the development of a new algorithm, termed mzCompare, to discover pure elution profiles for overlapped analytes in a GC-MS chromatogram. This algorithm identifies selective mass channels (*m/z*) for each analyte within a time window based on their similarity in t_R and peak shape. By discovering selective *m/z* for an analyte, their chromatographic signals are summed together to generate a pure elution profile. The pure elution profiles generated by mzCompare can then be used as a constraint in MCR-ALS to resolve both rotational ambiguities and the chemometric multiplex disadvantage. Furthermore, we also introduce the concept of a "resolved component" chromatogram to visualize the results obtained from the mzCompare algorithm. Using both experimentally collected and simulated GC-MS data, we will demonstrate the capability of mzCompare as a synergistic tool for MCR-ALS and its improvement on analyte identification and quantitation, especially at low R_s.

9.2. Methods and Materials

9.2.1. Method fundamentals

The development and application of the mzCompare algorithm was performed in Matlab 2019b (Mathworks, Inc., Natick, MA, USA). A peak finding algorithm, referred to as the

enhanced total ion current (ETIC) chromatogram, first moves through the baseline corrected GC-MS chromatogram to discover peak maxima on each *m/z* with signal greater than a user-defined signal-to-noise (*S/N*) threshold [33]. The *S/N* threshold used herein was 10. A tile window centered on each peak location is drawn and a t_R for every *m/z* discovered by the ETIC algorithm is measured. The width of the tile is designed to be slightly larger than the peak width-at-base (W_b). Note, at this stage the peak finder is locating all detectable peaks, whether they are singlets (pure peaks), or overlapped peaks (to be handled by mzCompare prior to MCR-ALS). Within this tile, an intra-chromatogram comparison of peak shape is performed, where a lack-of-fit (*LOF*) is calculated between every *m/*z passing the *S/N* threshold,

$$LOF = \sqrt{\frac{\sum_i (y-x)_i^2}{\sum_i x_i^2}} \times 100 \tag{9.1}$$

where *x* and *y* are the elution profiles for each *m/z* [34]. Ideally, *m/z* selective for the same analyte will have similar t_R and a low *LOF* while *m/z* belonging to different analytes will have different t_R and a large *LOF*. The t_R and *LOF* from these intra-chromatogram comparisons are visualized on a cluster plot. This plot is then reduced to only show the comparisons with a $LOF \leq$ 5 %, which is a user-defined threshold that signifies a comparison involving pure, selective *m/z* for an analyte [22,23,35]. From this reduced cluster plot, locations of pure analyte clusters are found using a density-based spatial clustering algorithm. Pure elution profile for each analyte cluster can be generated by summing together all the selective *m/z* discovered by mzCompare and these profiles can be used as an equality constraint in MCR-ALS. Furthermore, a "resolved component" chromatogram can also be created to place the information shown on the cluster plot back into chromatographic context. This resolved component chromatogram is generated by summing then representing the total peak area from the selective *m/z* discovered in each analyte cluster into a Gaussian profile with a W_b equal to the cluster width in the t_R dimension ($W_{b,cluster}$).

9.2.2. Experimental chromatograms

The experimental chromatograms presented herein were collected using GC instruments coupled to a time-of-flight mass spectrometer (GC-TOFMS). The chromatographic conditions used for each separation are provided in Appendix G but are described briefly herein. The first separation to be presented in this work is that of a 73-component (Table G.1) test mixture, where the TOFMS collected *m/z* 40 – 230 at 100 Hz [36]. Using a low thermal mass (LTM) resistive heating module, mzCompare is also applied to the separation of 115-component (Table G.2) test mixture [28]. Note, *m/z* 33 – 250 were collected at 500 Hz for the LTM-GC-TOFMS separations [28]. However, despite the fast collection frequency employed for this separation, the greater precision provided in the time domain made it unlikely that *m/z* from the same analyte would share a similar t_R [28]. Therefore, the time domain of the LTM-GC-TOFMS separation was boxcar averaged to 100 Hz. Lastly, mzCompare was also applied to the GC-TOFMS separation of an aerospace fuel, where *m/z* 45 – 300 were collected at 100 Hz.

MCR-ALS was applied to overlapped regions of the chromatograms prior to mzCompare using all *m/z*, a non-negativity constraint for both the chromatographic profiles and mass spectra, and a unimodality constraint for the chromatographic profiles. A forward match value (MV) was calculated by comparing the MCR-ALS resolved mass spectrum to the NIST 11 MS library (National Institute of Standards and Technology, Gaithersburg, MD, USA). A sufficient MS library match was declared with a MV ≥ 800 [37]. To determine the contribution of rotational ambiguity in MCR-ALS models, the MCR-BANDS algorithm developed by Jaumot and Tauler was also employed [38]. The mzCompare algorithm described earlier was then applied to these three chromatograms using a tile size of 2 s, 1 s, and 4s, respectively. The pure elution profiles developed by mzCompare were then used as a concentration equality constraint in successive

MCR-ALS models (i.e., mzCompare assisted MCR-ALS) developed for these chromatographic regions of interest.

9.2.3. Simulated chromatograms

GC-MS simulations of target-interferent pairs were generated based on the methodology provided in a previous report [21]. Each simulation consisted of two analytes, where one analyte was assigned as the target and the other was assigned as the interferent. A library of 45 analytes (Table G.3) was developed by randomly extracting mass spectra from the NIST 11 MS database. This library resulted in 990 target-interferent pairs, where the mass spectra similarity between a pair was defined according to their MV. Analyte pairs with a MV between 600 – 1000 were defined as having high mass spectra similarity while analyte pairs with a MV range of 300 – 600 and 0 – 300 were defined as having mid and low similarity, respectively. Each analyte was simulated as a Gaussian profile, with a W_b of 1 s and peak area of 10,000. The data collection rate was simulated at 100 Hz. Random Gaussian noise was added to every *m/z* in the simulation such that the *S/N* in the TIC chromatogram of each analyte equaled either 10 or 50. Sixteen R_s values were ultimately simulated (R_s = 0.02, 0.04, 0.08, 0.10, 0.12, 0.14, 0.16, 0.18, 0.20, 0.25, 0.30, 0.35, 0.40, 0.45, and 0.50). This procedure resulted in a total of 31,680 simulations (990 target-interferent pairs, 16 R_s values, and 2 *S/N* options).

Two-component MCR-ALS models were developed for every GC-MS simulation. These initial MCR-ALS models used the same non-negativity and unimodality constraints as described earlier. The mzCompare algorithm was then applied to these simulations using a tile size of 2.5 s. The pure elution profiles for the target and interferent analytes were then used as equality constraints in the mzCompare assisted MCR-ALS models for each simulation. Both the initial and mzCompare assisted MCR-ALS models were evaluated for their accuracy in identification

and quantitation. For identification accuracy, the MV between the resolved mass spectrum and the original library mass spectrum was calculated. To assess quantitation accuracy, the peak area obtained from the resolved elution profile (A_{model}) was compared to the simulated peak area (A_{sim}) using a percent error calculation [21]:

$$\%\, error = \frac{|A_{model} - A_{sim}|}{A_{sim}} \times 100 \tag{9.2}$$

9.3. Results and Discussion

The mzCompare algorithm was first applied to a separation of a 73-component test mixture collected with GC-TOFMS. From the TIC chromatogram of the mixture, a total of 64 peaks were detected (Figure 9.1A), whereby the W_b in the ranged from ~ 1.0 s to 1.2 s. In contrast, the results of the mzCompare algorithm are shown as a "resolved component" chromatogram in Figure 9.1B, where all 73 analytes in the test mixture are successfully resolved. This visualization method simulates each analyte resolved by mzCompare as a Gaussian profile with a new W_b equal to the width of its analyte cluster (i.e., $W_{b,\text{cluster}}$), where the $W_{b,\text{cluster}}$ ranged from 20 ms to 40 ms. A more detailed explanation of the definition of $W_{b,\text{cluster}}$ is forthcoming, *vide infra*. The improved peak detection provided by mzCompare is due to its unique ability to discover pure, selective *m/z* that belong to overlapped analytes using a single chromatogram. Figure 9.1C-D provides a zoom-in on a complex region of the chromatogram just before 4 min to illustrate the differences in W_b between the original TIC chromatogram (C) and resolved component chromatogram (D). Since the mzCompare algorithm reduces the original chromatogram into pure, resolved analyte *m/z* clusters, the R_s between overlapped species and overall peak capacity (n_c) of the separation increases. For the separation shown in Figure 9.1A,C, the standard definition of R_s and n_c are:

$$R_S = \frac{2(t_{R_2} - t_{R_1})}{W_{b_2} + W_{b_1}} \tag{9.3}$$

$$n_c = \frac{t_{sep}}{\overline{W}_b}(R_s = 1) \quad (9.4)$$

For the resolved component chromatogram shown in Figure 9.1B,D, the $R_{s,mzCompare}$ and $n_{c,mzCompare}$ can be expressed by replacing W_b in Eqs. 9.3-9.4 with $W_{b,cluster}$. As a result, the enhancement in R_s and n_c provided by mzCompare is:

$$\frac{R_{s,mzCompare}}{R_s} = \frac{W_{b_2}+W_{b_1}}{W_{b,cluster_2}+W_{b,cluster_1}} \quad (9.5)$$

$$\frac{n_{c,mzCompare}}{n_c} = \frac{\overline{W}_b}{\overline{W}_{b,cluster}} \quad (9.6)$$

Therefore, using the average W_b and $W_{b,cluster}$ measured in Figure 9.1A-B as inputs to Eqs. 9.5-9.6, roughly a 40-fold increase in R_s and n_c is provided by mzCompare.

To demonstrate how the original chromatogram in Figure 9.1A,C is translated into the resolved component chromatogram in Figure 9.1B,D, the entire workflow of the mzCompare algorithm will first be illustrated on the chromatographic peak highlighted by the yellow star. Figure 9.1E-F shows all the *m/z* signals overlaid for this chromatographic region. The unresolved analyte pair shown here is 1-octanol and butylbenzene. Based on a t_R difference of 100 ms and a W_b of 1.15 s, the R_s between 1-octanol and butylbenzene is 0.1. Due to the poor R_s between these two analytes, only one peak can be detected in the TIC chromatogram. However, to improve analyte detection, the newly developed mzCompare algorithm will be applied to objectively discover the pure, selective *m/z* for each overlapped analyte. The information provided by mzCompare will not only be used to demonstrate R_s enhancement, but we will also show how this algorithm can be a synergistic tool to improve MCR-ALS identification.

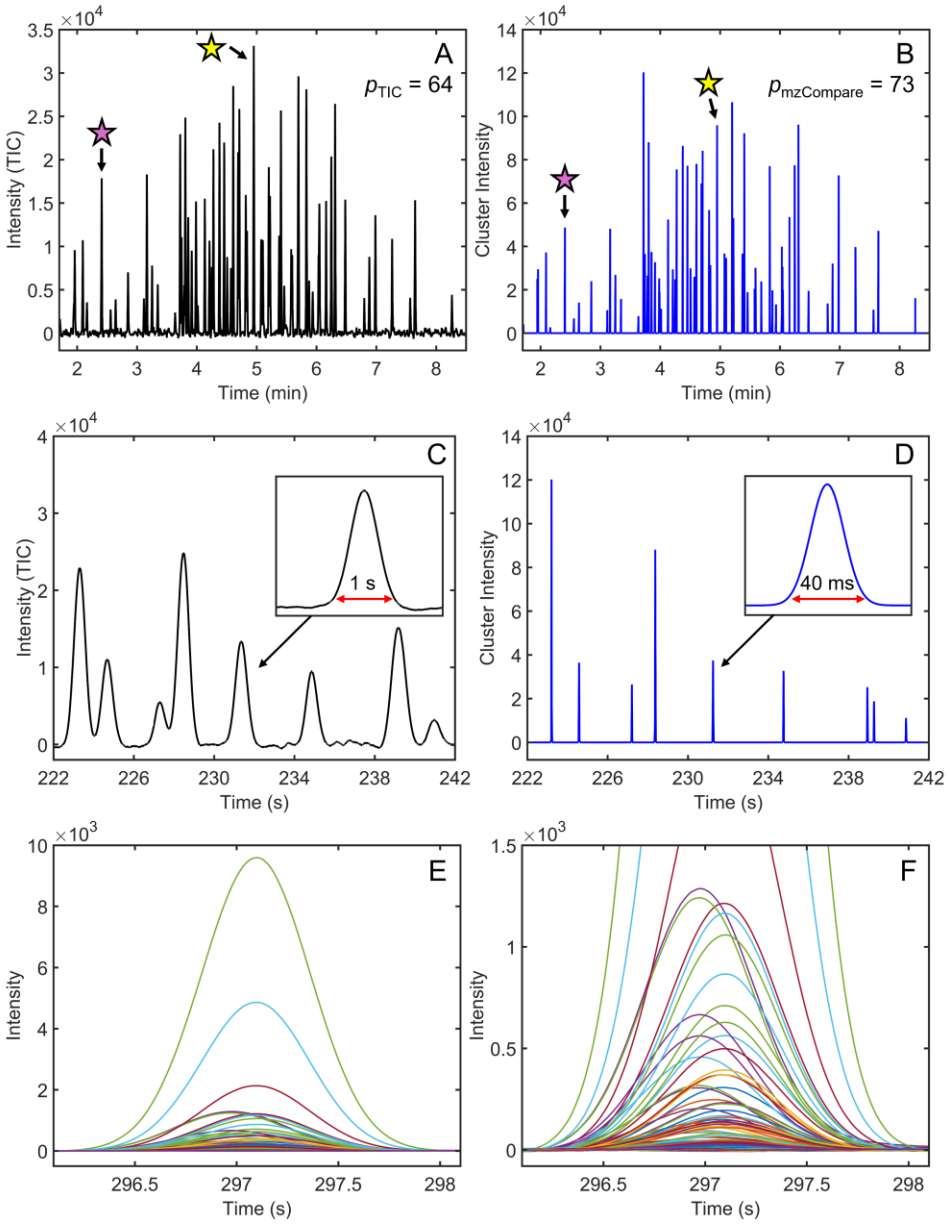


Figure 9.1. (A) Total ion current (TIC) chromatogram of the test mixture of a 73-component test mixture separated using GC-TOFMS. The number of observed peaks (p) is provided. Two chromatographic regions of interest for subsequent examination are labeled by a yellow and pink star, respectively. (B) The resolved component chromatogram generated for the separation in (A) after applying the mzCompare algorithm. The total number of observed peaks (p) and two peaks of interest (yellow and pink stars) are labeled. (C) A zoom-in on a highly overlapped chromatographic region in (A). Inset: Demonstration of the typical peak width (W_b) in the chromatogram, where W_b is 1 s. The x-axis scale is 229.5 – 233 s and y-axis scale is -1000 – 15000. (D) The resolved component chromatogram for the region in (C). Inset: Demonstration of the cluster peak width ($W_{b,cluster}$), where $W_{b,cluster}$ is 40 ms. The x-axis scale is 231.2 – 231.1 s and y-axis scale is -3000 – 40000. (E) The chromatographic peak of interest labeled by the yellow star in (A) with the signal from all m/z provided. This peak is made up of two overlapped analytes: 1-octanol and butylbenzene. (F) A zoom-in on the peak shown in (E).

The mzCompare algorithm discovers pure, selective *m/z* for each analyte in the chromatogram by implementing an intra-chromatogram peak shape comparison. Using a single chromatogram, peak shape comparisons in the mzCompare algorithm are performed by calculating the *LOF* between all *m/z* above the specified *S/N* threshold. Figure 9.2 illustrates the intra-chromatogram *LOF* comparisons to determine *m/z* purity. The overlaid chromatograms at *m/z* 55 (red) and *m/z* 68 (blue) in Figure 9.2A demonstrate that these two *m/z* share a similar t_R (297.11 s and 297.10 s). After normalizing the signals to the same peak area, low residuals (black) between the *m/z* 55 (red) and *m/z* 68 (blue) are obtained along with a small *LOF* of 4.7 % (Figure 9.2B). Based on the similarity in t_R and low *LOF* of 4.7 %, it was determined that these two *m/z* are pure for 1-octanol. Figure 9.2C shows the chromatograms at *m/z* 55 (red) and *m/z* 77 (blue), which have respective t_R of 297.11 s and 297.20 s. Due to the difference in t_R and peak shapes, high residuals (black) were calculated between the normalized chromatograms, resulting in a *LOF* of 65 % (Figure 9.2D). Hence, these *m/z* are not shared by the same analyte. Since it was determined that 1-octanol did not contribute to the signal at *m/z* 77, Figure 2E shows the overlay of this *m/z* (red) to *m/z* 78 (blue). Both *m/z* share the same t_R of 297.20 s. Following the area normalization of these two chromatograms, low residuals (black) and *LOF* were calculated (Figure 9.2F). Based on their similar t_R and low *LOF* of 1.5 %, it was determined that these two *m/z* are pure for butylbenzene. Ultimately, the identification of these pure analyte *m/z* through this intra-chromatogram *LOF* comparison enables mzCompare to resolve the individual analytes within a given peak.

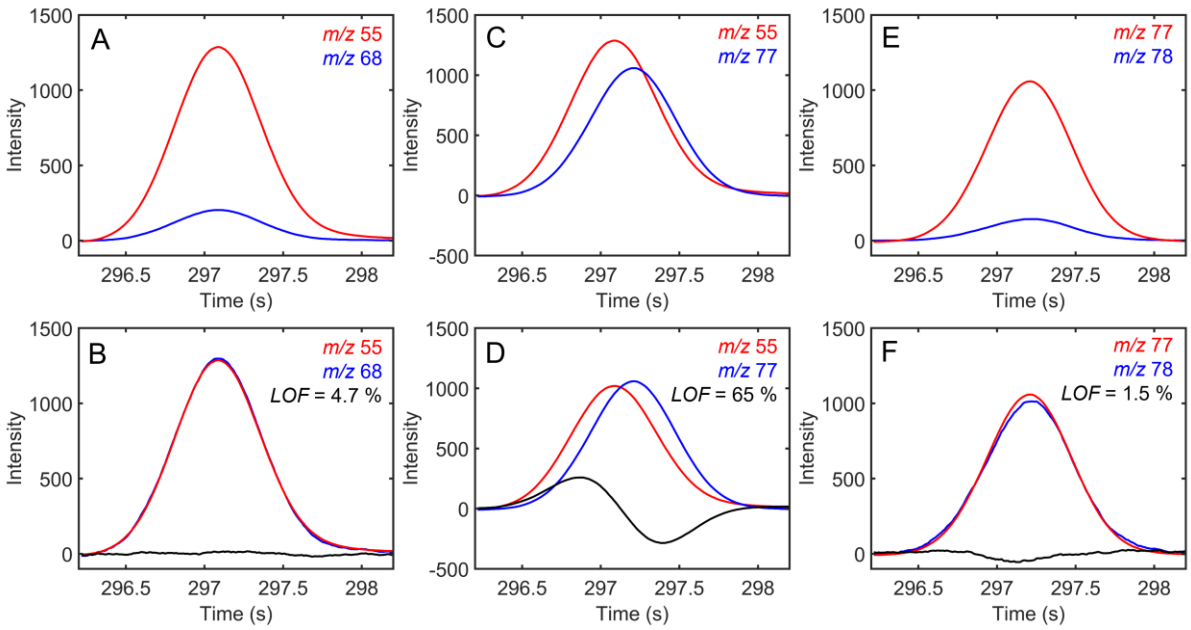


Figure 9.2. Demonstration of the intra-chromatogram lack-of-fit (*LOF*) calculation for the separation of 1-octanol and butylbenzene. (A) An overlay of the signal on *m/z* 55 (red) and *m/z* 68 (blue). (B) An overlay of the normalized signals in (A) along with the *LOF* residuals. (C) An overlay of the signal on *m/z* 55 (red) and *m/z* 77 (blue). (D) An overlay of the normalized signals in (C) along with the *LOF* residuals. (E) An overlay of the signal on *m/z* 77 (red) and *m/z* 78 (blue). (F) An overlay of the normalized signals in (E) along with the *LOF* residuals.

The next step in the mzCompare algorithm is to reduce these intra-chromatogram *LOF* comparisons down into a cluster plot, which in turn ultimately provides the $W_{b,cluster}$ for each peak in the resolved component chromatogram. Figure 9.3A illustrates the *LOF* calculated between every *m/z* pair for the separation of 1-octanol and butylbenzene, where *m/z* comparisons that had a *LOF* < 20 % are shaded blue and comparisons that has a *LOF* > 20 % are shaded black. Note, Table 9.1 provides an index key for these *m/z* along with their t_R and analyte identity. Based on Figure 9.3A, intra-chromatogram comparisons with *m/z* from different analytes had *LOF*s > 20 % due to their dissimilar t_Rs and peak shapes. Hence, using only the *m/z* comparisons with a *LOF* < 20 %, a cluster plot can be developed for this separation (Figure 9.3B). Each bin on the cluster plot in Figure 9.3B has a bin size of 10 ms × 1 % *LOF* and these bins are color coded based on their occurrence frequency (i.e., how many *m/z* comparisons are in

each bin on the plot). Previous work with comparison-based studies involving 2 or more chromatograms deemed that a $LOF \leq 5$ % was sufficient at discovering pure *m/z* for a target analyte, which have little to no signal contributions from interferent analyte(s) [22,23,35]. Therefore, a *LOF* threshold of 5 % was utilized in the mzCompare methodology to discover the intra-chromatogram comparisons involving pure, selective *m/z* for 1-octanol and butylbenzene. After application of this threshold, a density-based spatial clustering algorithm discovered the location of two analyte clusters, which were centered around a t_R 297.10 s or 297.20 s (black dashed boxes on Figure 9.3B). It is also important to note that this cluster plot has a "horseshoe-like" shape, where intermediate t_R (297.13 – 297.15 s) have larger *LOF*s (*LOF* > 13 %) compared to the *m/z* comparisons centered within the pure analyte cluster boxes. The bins at the top of this "horseshoe" shape (within in the gray dotted box on Figure 9.3B) represent *m/z* comparisons involving either *m/z* 41 or *m/z* 53, which are shared by 1-octanol and butylbenzene (Table 9.1). Hence, the *LOF* for *m/z* comparisons involving either one of these shared *m/z* is between 5 % and 20 % because both analytes contribute to the net signal observed.

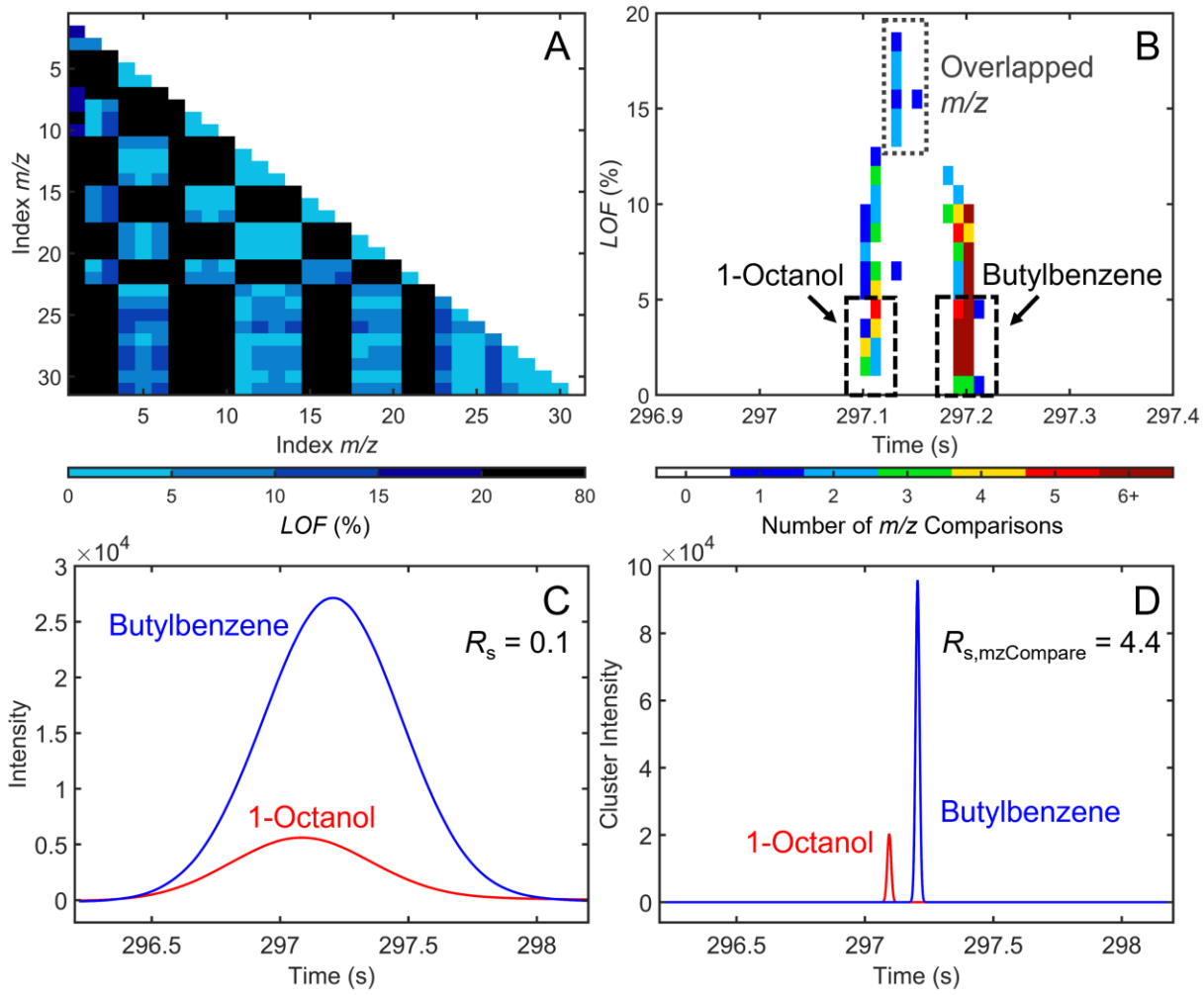


Figure 9.3. Application of mzCompare to the for the separation of 1-octanol and butylbenzene. (A) The intra-chromatogram *LOF*s determined for each *m/z* pair (see Table 9.1 for the *m/z* index key). (B) Cluster plot of intra-chromatogram *LOF* versus retention time (t_R) for *m/z* comparisons in (A) that had a *LOF* < 20 %. The bins are color coded according to the frequency of their occurrence. The black dashed boxes represent the pure analyte clusters for 1-octanol and butylbenzene. The dotted gray box represents the *LOF* comparisons involving a *m/z* shared by both analytes. (C) Overlay of the analytical ion current (AIC) chromatogram generated for 1-octanol (red) and butylbenzene (blue) using the pure *m/z* discovered in (B). The original R_s was 0.1. (D) The resolved component chromatogram for 1-octanol (red) and butylbenzene (blue). The new $R_{s,mzCompare}$ equals 4.4. Note, the resolved component chromatogram shown here correlates to a zoom-in on the region marked by the yellow star in Figure 9.1B.

Table 9.1. The index key for the application of mzCompare to the peak containing 1-octanol and butylbenzene, where *m/z* with an intensity above the minimum threshold were considered for the *LOF* comparisons. The retention time (t_R) measured for each *m/z* is provided along with the analyte(s) that each *m/z* belongs to.

Index Value	*m/z*	t_R (s)	Analyte Responsible for *m/z*
1	41	297.13	Shared
2	42	297.11	1-Octanol
3	43	297.11	1-Octanol
4	50	297.20	Butylbenzene
5	51	297.19	Butylbenzene
6	52	297.20	Butylbenzene
7	53	297.15	Shared
8	55	297.11	1-Octanol
9	56	297.10	1-Octanol
10	57	297.11	1-Octanol
11	62	297.19	Butylbenzene
12	63	297.20	Butylbenzene
13	64	297.20	Butylbenzene
14	65	297.20	Butylbenzene
15	68	297.10	1-Octanol
16	69	297.10	1-Octanol
17	70	297.11	1-Octanol
18	77	297.20	Butylbenzene
19	78	297.20	Butylbenzene
20	79	297.20	Butylbenzene
21	83	297.10	1-Octanol
22	84	297.10	1-Octanol
23	89	297.19	Butylbenzene
24	91	297.20	Butylbenzene
25	92	297.19	Butylbenzene
26	93	297.18	Butylbenzene
27	103	297.20	Butylbenzene
28	104	297.19	Butylbenzene
29	105	297.21	Butylbenzene
30	115	297.20	Butylbenzene
31	134	297.20	Butylbenzene

The analyte clusters (black dashed boxes) shown in Figure 9.3B define all the selective *m/z* required to generate a purified peak profile for 1-octanol and butylbenzene. Since the mass spectra were collected at 100 Hz, each pixel in the t_R domain equals 10 ms. As shown in Figure 9.3, the $W_{b,cluster}$ for 1-octanol and butylbenzene are 20 ms (2 pixels wide) and 30 ms (3 pixels

wide), respectively. Table 9.1 defines which *m/z* were found to be pure for 1-octanol or butylbenzene. Figure 9.3C shows the overlaid analytical ion current (AIC) chromatograms for 1-octanol (red) and butylbenzene (blue), which were generated by summing together the signal from the pure analyte *m/z* discovered in Figure 9.3B and Table 9.1. These AIC chromatograms also illustrate the original R_s between the 1-octanol and butylbenzene was 0.1 (Figure 9.1E-F). To further visualize the mzCompare results obtained in Figure 9.3B-C, the resolved component chromatogram for 1-octanol (red) and butylbenzene (blue) can be generated (Figure 9.3D). Note, Figure 9.3D is also a zoom-in on the region marked by a yellow star in Figure 9.1B. The resolved component chromatogram combines the information in Figure 9.3B-C, where new peak profiles are simulated using the $W_{b,cluster}$ (Figure 9.3B) and peak area equal to the area for each peak the AIC (Figure 9.3C). Based on the resolved component chromatogram in Figure 9.3D, the mzCompare methodology increases the R_s between these two analytes ($R_{s,mzCompare}$) to 4.4. Thus, the mzCompare algorithm provided a 44-fold increase in R_s between 1-octanol and butylbenzene (Eq. 9.5), improving the detection of both analytes.

While the resolved component chromatogram generated by the mzCompare method can improve analyte detection, the selective *m/z* discovered herein can also be used to enhance analyte identification. Chemometric decomposition methods like MCR-ALS are typically used to obtain the purified chromatographic profiles and *m/z* signals for each overlapped analyte. However, achieving accurate analyte identifications with chemometric decomposition can be challenging for analytes with a low R_s, low *S/N*, and/or a high degree of spectrum contamination [19–23,39,40]. For example, the separation of 1-octanol and butylbenzene can be a challenge to chemometric decomposition efforts because of its low R_s of 0.1. Figure 9.4A-B shows the MCR-ALS resolved mass spectrum for 1-octanol (A; red) and butylbenzene (B; blue) compared against

their library spectra. A MV of 288 was calculated for 1-octanol (Figure 9.4A) while a MV of 910 was calculated for butylbenzene (Figure 9.4B). With a MV ≥ 800 often required to declare a sufficient match [37], only butylbenzene could be confidently identified in this chromatographically overlapped peak. Previous work has attributed the poor performance of an MCR-ALS model to rotational ambiguity, a type of uncertainty that occurs when there are multiple feasible solutions that equally fit the model [38,41,42]. The MCR-BANDS algorithm was developed to evaluate the presence of rotational ambiguities in a model by determining the minimum and maximum relative contributions for each component [38,41,42]. If rotational ambiguities are present in the model, then a range of possible MV for each analyte can be obtained from the minimum and maximum spectral contributions. Application of MCR-BANDS to the model in Figure 9.4A-B illustrated that rotational ambiguities are present in the model, with 1-octanol and butylbenzene having a respective MV range between 230 – 300 and 904 – 915. However, the presence of rotational ambiguities in this model cannot solely explain why MCR-ALS was incapable of resolving a pure mass spectrum for 1-octanol to achieve a MV ≥ 800 for analyte identification. Visual examination of Figure 9.4A-B demonstrates that *m/z* signals from butylbenzene are present in the MCR-ALS resolved mass spectrum for 1-octanol. The presence of these interfering *m/z* signals is due to the "chemometric multiplex disadvantage" [22,23]. This disadvantage occurs when the decomposition model includes noise and signal from larger interfering compounds in the resolved mass spectrum of the target analyte as a means of minimizing the residual error. In other words, the MCR-ALS model is unable to resolve 1-octanol because its efforts are dominated by larger interfering signals from butylbenzene.

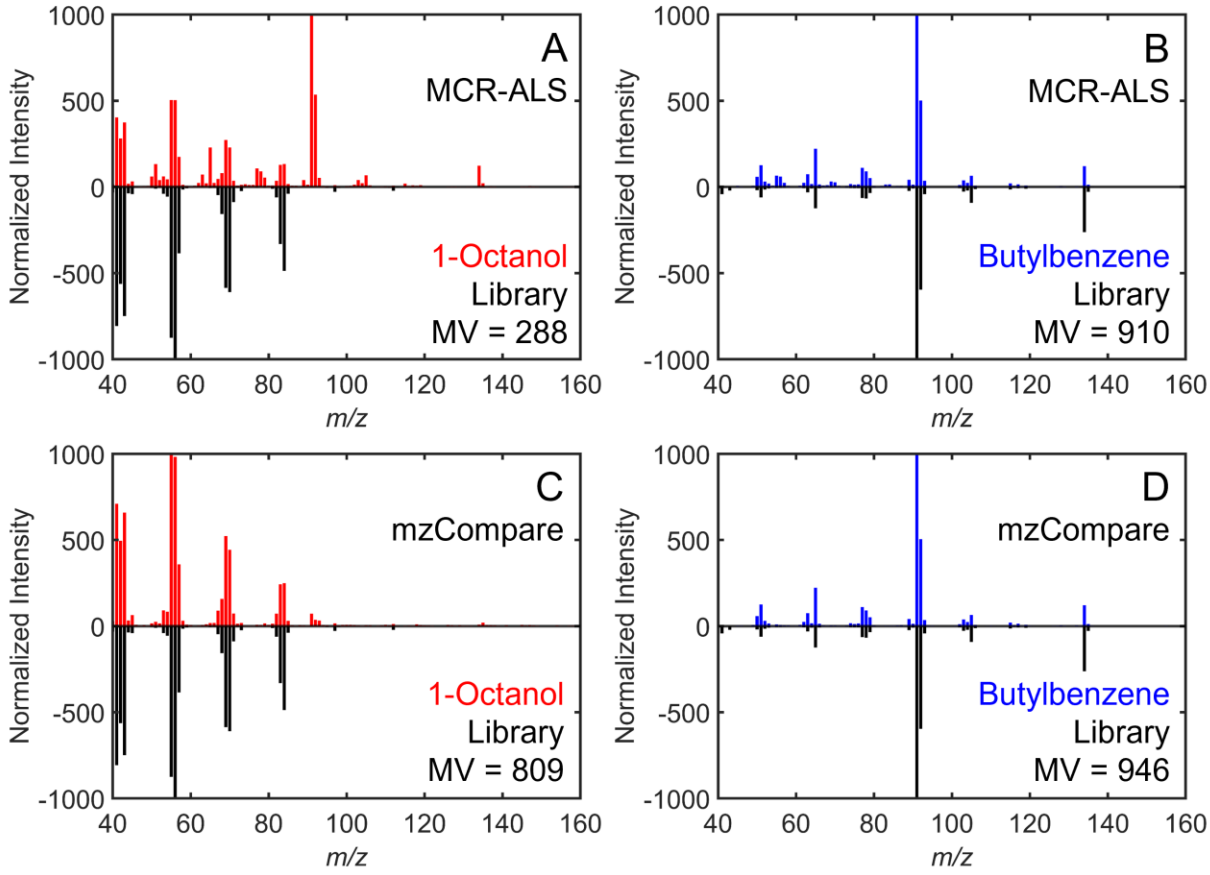


Figure 9.4. Improvement in peak identification with mzCompare assisted MCR-ALS. (A) Reflection plot of the initial MCR-ALS spectrum for 1-octanol (red) and its library spectrum (black). (B) Reflection plot of the initial MCR-ALS spectrum for butylbenzene (blue) and its library spectrum (black). (C) Reflection plot of the spectrum obtained for 1-octanol (red) from mzCompare assisted MCR-ALS and its library spectrum (black). (D) Reflection plot of the spectrum obtained for butylbenzene (blue) from mzCompare assisted MCR-ALS and its library spectrum (black). For all panels, a match value (MV) is provided.

To further illustrate the promise of mzCompare for analyte detection and identification, this algorithm was then applied to the very challenging situation provided in the separation of benzene and cyclohexane. Figure 9.5A-B shows a zoom-in on the chromatographic region highlighted by the pink star in Figure 9.1A, with every *m/z* overlaid. The t_R and W_b measured for benzene was 144.45 s and 1.11 s, respectively, while the corresponding measurements for cyclohexane was 144.51 s and 1.14 s. Therefore, the original R_s between these two analytes equaled 0.05. Figure 9.5C-D shows the resulting mass spectra obtained for benzene (C; red) and

cyclohexane (D; blue) based on the initial MCR-ALS decomposition of this chromatographic time window. The MV obtained for the MCR-ALS resolved mass spectrum of benzene was 904, meeting the threshold for confident identification (Figure 9.5C). However, Figure 9.5D reveals that the *m/z* signals for benzene are contaminating the decomposed mass spectrum for cyclohexane, resulting in a poor MV of 435. MCR-BANDS indicated that the minimum and maximum spectral contribution for benzene would cause the MV to range between 893 – 912 while the MV range for cyclohexane was 432 – 464. Hence, the MCR-ALS model in Figure 9.5C-D is affected by both rotational ambiguities and the chemometric multiplex disadvantage. Application of mzCompare is necessary to improve both the detection and identification of these two analytes.

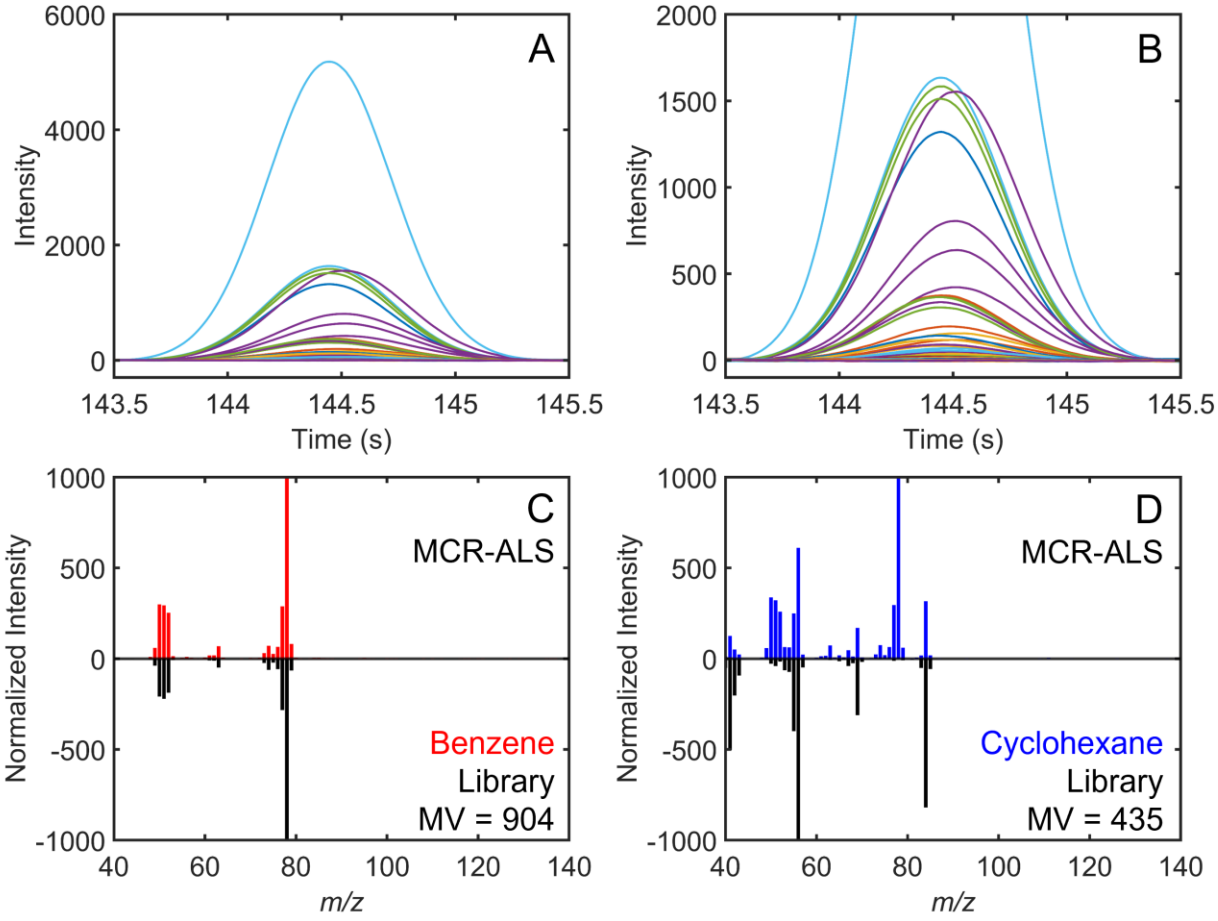


Figure 9.5. Illustration of the challenge associated with identifying benzene and cyclohexane (R_s = 0.05). (A) The chromatographic peak of interest labeled by the pink star in Figure 9.1A with the signal from all *m/z* provided. (B) A zoom-in on the peak shown in (A). (C) Reflection plot of the initial MCR-ALS spectrum for benzene and its library spectrum (black). (D) Reflection plot of the initial MCR-ALS spectrum for cyclohexane (blue) and its library spectrum (black). A MV is provided for both (C-D).

Figure 9.6 shows the application of mzCompare to the chromatographic time window containing benzene and cyclohexane with a R_s of 0.05. From the plot shown in Figure 9.6A, mzCompare discovered two pure analyte clusters (dashed black boxes), representing all the selective *m/z* for benzene (t_R = 144.45 s) and cyclohexane (t_R = 144.51 s). Again, a "horseshoe-like" shape can be observed on the cluster plot due to the shared *m/z* between these two analytes (*m/z* 50 and *m/z* 51). However, it is important to note that any comparison involving these two shared *m/z* still had a *LOF* > 5 % despite the low R_s between these two analytes. Based on the

pure analyte clusters in Figure 9.6A, the $W_{b,cluster}$ for benzene and cyclohexane are both 30 ms. Figure 9.6B shows the resolved component chromatogram for benzene (red) and cyclohexane (blue) that was generated using the pure analyte *m/z* discovered by mzCompare. Note, Figure 9.6B is also a zoom-in of the chromatographic window highlighted by the pink star in Figure 9.1B. The $R_{s,mzCompare}$ between these two analytes in the resolved component chromatogram is 2.0 (a 40-fold improvement). Due to this enhancement in R_s, both benzene and cyclohexane are now detectable as individual peaks. The inset in Figure 9.6B provides an overlay of the AIC chromatograms developed for these two analytes, further demonstrating the original R_s of 0.05. The pure chromatographic profiles in the AIC were then used as equality constraints in a consecutive MCR-ALS model (Figure 9.6C-D). After mzCompare assisted MCR-ALS, both analytes can be confidently identified with a MV of 921 for benzene (Figure 9.6C) and a MV of 885 for cyclohexane (Figure 9.6D). The results from MCR-BANDS also highlighted that this new model was not affected by rotational ambiguity. Therefore, even under these challenging chromatographic conditions, the pure elution profiles provided by mzCompare enabled MCR-ALS to overcome the effects of rotational ambiguity and the chemometric multiplex disadvantage.

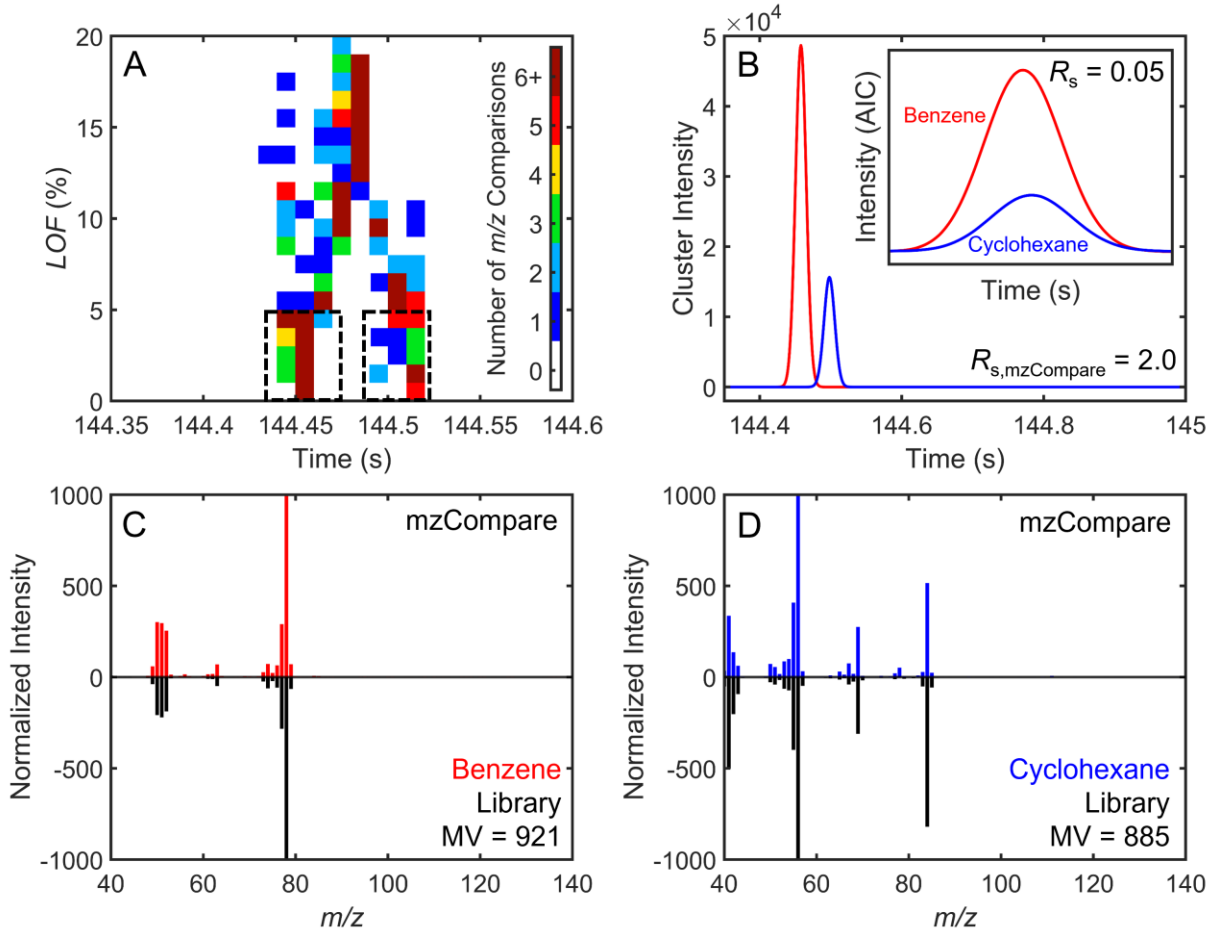


Figure 9.6. Application of mzCompare to the separation of benzene and cyclohexane. (A) Cluster plot of intra-chromatogram *LOF* calculations versus t_R for benzene and cyclohexane. Note, only comparisons with a *LOF* < 20 % are shown. The black dashed boxes represent the pure analyte clusters for benzene and cyclohexane. (B) The resolved component chromatogram for benzene (red) and cyclohexane (blue), which corresponds to a zoom-in on the region highlighted by the pink star in Fig. 1B. The $R_{s,mzCompare}$ equals 2.0. Inset: The AIC chromatogram generated for benzene (red) and cyclohexane (blue) using the pure *m/z* discovered in (A). The x-axis scale is 143.5 – 145.5 s and the y-axis scale is -1000 – 15000. The original R_s between these two analytes was 0.05. (C) Reflection plot of the spectrum obtained for benzene (red) from mzCompare assisted MCR-ALS and its library spectrum (black). (D) Reflection plot of the spectrum obtained for cyclohexane (blue) from mzCompare assisted MCR-ALS and its library spectrum (black). A MV is provided for both (C-D).

Based on the promising application of mzCompare to the 73-component test mixture shown in Figure 9.1A-B, the performance of this new algorithm was then tested using a separation of 115-component test mixture collected with LTM-GC-TOFMS (Figure 9.7). The top row shows the (A) TIC chromatogram before mzCompare and (B) resolved component

chromatograms after mzCompare. The bottom row provides a (C) zoom-in on a highly overlapped region in the TIC chromatogram and (D) view of that same region in the resolved component chromatogram. Compared to a traditional air-bath GC instrument, the LTM-GC platform can produce fast separations with narrow W_b through indirect resistive heating of the column [28,43]. Figure 9.7A illustrates this idea, where the total separation window is ~ 120 s long and the W_b range from 300 to 500 ms. However, despite the narrow W_b produced, the speed of the separation causes many analytes to co-elute. As a result, only 83 peaks were detected in the TIC chromatogram (Figure 9.7A). Fortunately, the application of mzCompare provided a 39 % increase in the number of peaks detected, by discovering all 115 analytes in the test mixture (Figure 9.7B). Figure 9.7C-D demonstrates the power of mzCompare by resolving 12 analytes in a 7 s time window. The TIC chromatogram in Figure 9.7C demonstrates that only 7 peaks would be detected in this region and analyte identification would be challenging due to the low R_s. Whereas Figure 9.7D shows that mzCompare detected all 12 analytes as individual peaks with a $R_s > 1$. With a MV $\geq$ 800, these analytes were identified by mzCompare assisted MCR-ALS as: (1) chloroform, (2) 1-hexyne, (3) isobutanol, (4) methylcyclopentane, (5) *t*-amyl alcohol, (6) 1,1,1-trichloroethane, (7) 1-chlorobutane, (8) 1-butanol, (9) benzene, (10) neopentyl alcohol, (11) cyclohexane, and (12) carbon tetrachloride.

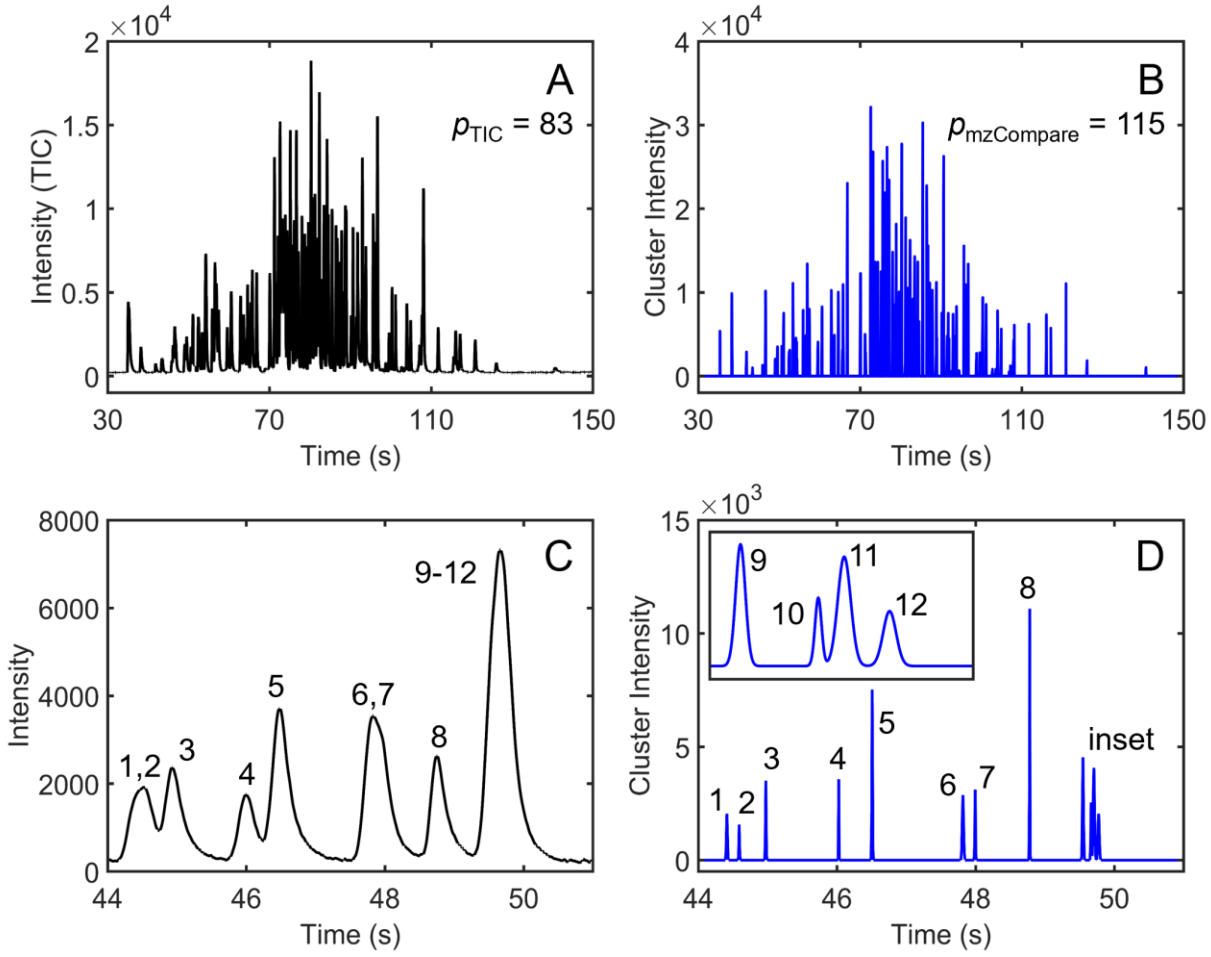


Figure 9.7. Application of mzCompare to the separation of a 115-component test mixture collected using low thermal mass (LTM)-GC-TOFMS. The (A) TIC and (B) resolved component chromatograms are provided along with the number of peaks (p) detected. A zoom-in on an unresolved chromatographic region in the (C) TIC and (D) resolved component chromatograms is also provided. Peaks identified in (C-D) are labeled as: (1) chloroform, (2) 1-hexyne, (3) isobutanol, (4) methylcyclopentane, (5) *t*-amyl alcohol, (6) 1,1,1-trichloroethane, (7) 1-chlorobutane, (8) 1-butanol, (9) benzene, (10) neopentyl alcohol, (11) cyclohexane, and (12) carbon tetrachloride. Inset: The x-axis scale is 49.5 – 49.9 s and the y-axis scale is -500 – 5000.

Peak detection and identification in aerospace fuels like Figure 9.8 can also be difficult because these samples are comprised of hundreds of chemical species with various functional groups. Yet standard one-dimensional GC methods do not have a sufficient peak capacity to resolve all these analytes within a reasonable run time [13]. Due to this limitation, only 233 peaks were detected in the TIC chromatogram of this aerospace fuel (Figure 9.8A). After

performing the mzCompare methodology, 335 peaks were detected (Figure 9.8B). The results in Figure 9.8A-B demonstrate a 44 % increase in the number of peaks detected after mzCompare. Figure 9.8C provides a zoom-in on the chromatographic region between 1152 s and 1171 s, where only 5 peaks can confidently be detected with a considerable *S/N*. For instance, peaks #1 and #2 in Figure 9.8C were undiscoverable with a typical peak detection algorithm because the signal from these analytes was lower than intensity of the background noise in the TIC chromatogram [33]. However, Figure 9.8D illustrates that mzCompare was able to detect 8 analytes within this separation window despite the low *S/N* for some analytes. The use of mzCompare assisted MCR-ALS was able to confidently identify (MV ≥ 800) these 8 analytes as: (1) butylcyclopentane, (2) 1,1-dimethylpropylbenzene, (3) 2,3-dimethyloctane, (4) 2,6-dimethyldecane, (5) 1-methyl-4(1-methylpropyl)benzene, (6) 1,2,4,5-tetramethylbenzene, (7) pentylbenzene, and (8) 2-methylundecane. In summary, Figure 9.7 and Figure 9.8 demonstrate that mzCompare can improve the analysis of complex sample mixtures due to the enhancement in R_s and n_c provided (Eqs. 9.5-9.6).

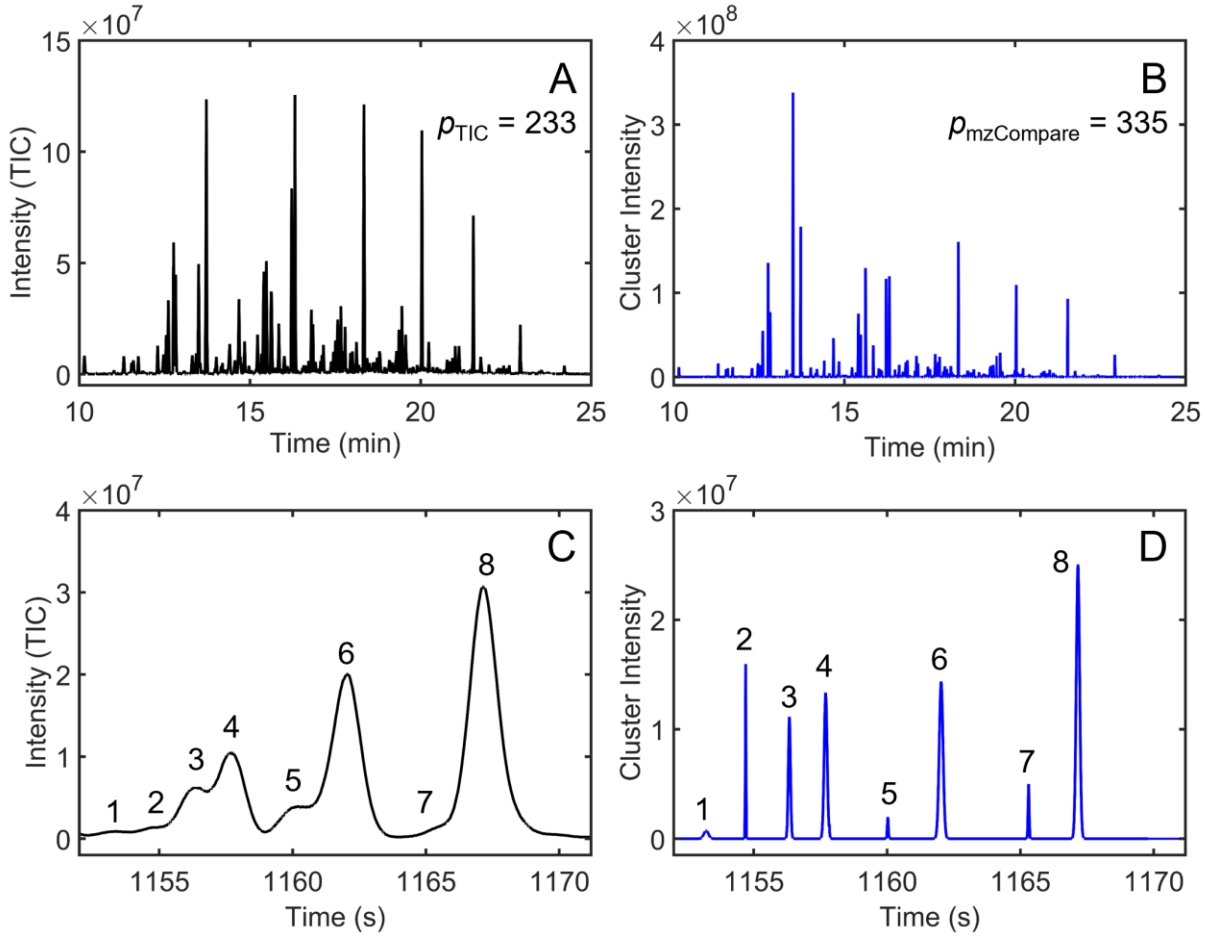


Figure 9.8. Application of mzCompare to the separation of an aerospace fuel collected with GC-TOFMS. The (A) TIC and (B) resolved component chromatograms are provided along with the number of peaks (*p*) detected. A zoom-in on an unresolved chromatographic region in the (C) TIC and (D) resolved component chromatograms is also provided. Peaks identified in (C-D) are labeled as: (1) butylcyclopentane, (2) 1,1-dimethylpropylbenzene, (3) 2,3-dimethyloctane, (4) 2,6-dimethyldecane, (5) 1-methyl-4(1-methylpropyl)benzene, (6) 1,2,4,5-tetramethylbenzene, (7) pentylbenzene, and (8) 2-methylundecane.

Lastly, chromatographic simulations of target-interferent pairs at R_s ranging from 0.02 to 0.5 were executed to demonstrate the performance of the mzCompare algorithm in overcoming the chemometric multiplex disadvantage. A total of 990 target-interferent pairs, with low (green), mid (blue) and high (red) degrees of mass spectral similarity, were simulated at every R_s (Figure 9.9A). A representative TIC chromatogram for an analyte simulated at an S/*N* of 50 and 10 is provided in Figure 9.9B. Figure 9.9C-D summarizes (C) the average MV and (D) the average

percent error in peak area (Eq. 9.2) obtained for the target analyte with the initial MCR-ALS model. The results for the target-interferent pairs are colored according to their mass spectral similarity (Figure 9.9A), with solid and dashed lines representing simulations at an *S/N* of 50 and 10, respectively. For all R_s, the initial MCR-ALS model resulted in a higher MV for target analytes simulated at an S/*N* of 50 compared to those at an *S/N* of 10 at all R_s (Figure 9.9C). The lower MVs obtained for the *S/N* 10 simulations is because the initial model included signals from noisy *m/z* into the decomposed mass spectrum for the target analyte. The inclusion of these noisy *m/z* along with signal from the interfering analytes increases as R_s decreases due to the chemometric multiplex disadvantage. Depending on the mass spectral similarity between the target and interferent analytes, the initial MCR-ALS models show that the lowest R_s required for confident target analyte identification (i.e., an average MV ≥ 800) was between 0.14 – 0.25 for a *S/N* of 10 and 0.02 – 0.1 for a *S/N* of 50 (Figure 9.9C). It is interesting to note that the target-interferent pairs with similar mass spectra have a higher average MV obtained for the target compared to the analyte pairs with low mass spectral similarity (Figure 9.9C). It would be expected that analyte pairs with similar mass spectra would pose a greater challenge for decomposition models, leading to lower target analyte MV. However, since these high similarity analyte pairs share many *m/z*, any signal from the interferent attributed to the mass spectrum of the target will not significantly impact its MV [21]. Whereas, when *m/z* signal from a dissimilar interferent is included into the target analyte mass spectrum, this will detrimentally affect the target MV [21]. This is not to say that the models developed for high similarity target-interferent pairs are not affected by the chemometric multiplex advantage. Instead, the effects from this disadvantage can be better observed in the quantitative information provided by the model, as described next.

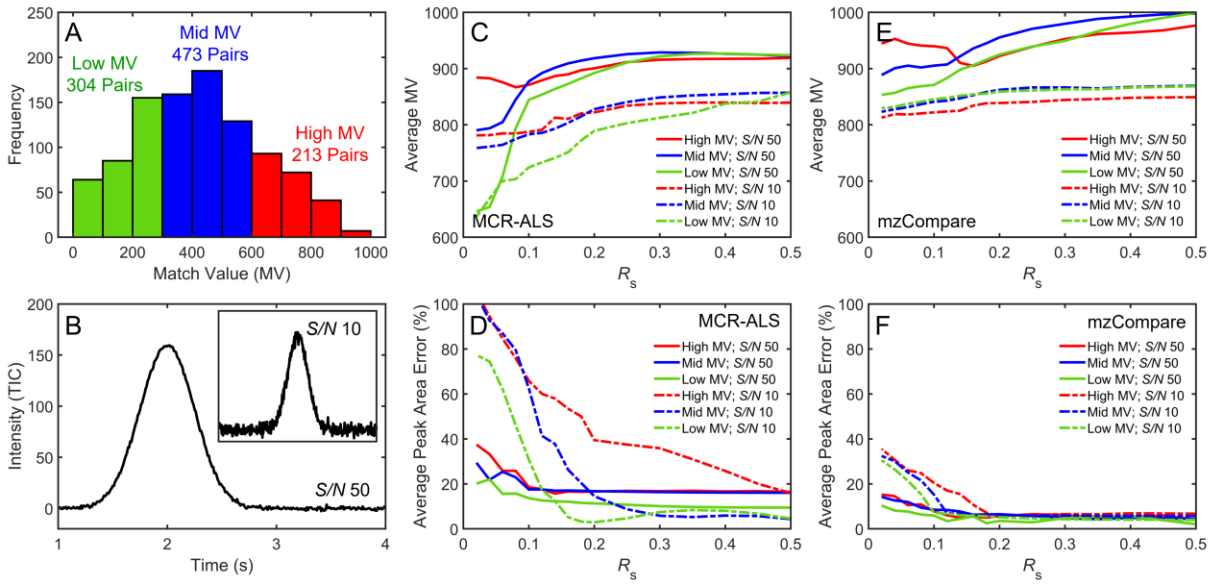


Figure 9.9. Summary of the model results versus R_s for target-interferent simulation study. (A) The MV calculated for the 990 target-interferent pair combinations. "Low" MV pairs (green) were identified as a MV < 300, "Mid" MV pairs (blue) had a MV between 300 – 600, and "High" MV pairs (red) had a MV > 600. (B) Simulated total ion current (TIC) chromatogram at a signal-to-noise ratio (*S/N*) of 50. Inset: Simulated TIC chromatogram at a *S/N* of 10. The x-axis scale is 0 – 4 s and y-axis scale is -20 – 200. (C) The average MV calculated between the mass spectrum extracted by the initial MCR-ALS model for the target analyte and the simulated mass spectrum. (D) The average quantitative error (Eq. 9.2) due to the difference in the peak area extracted by the initial MCR-ALS model for the target analyte and the peak area simulated. (E) The average MV calculated between the mass spectrum extracted by the mzCompare assisted MCR-ALS model and the simulated mass spectrum. (F) The average quantitative error due to the difference in the peak area extracted by the mzCompare assisted MCR-ALS model and the peak area simulated. Panels (C-F): The solid lines correspond to the simulations at a *S/N* of 50 while the dashed lines correspond to the simulations at a *S/N* of 10. The line colors correspond to the similarity between the target and interferent, as shown in (A).

Figure 9.9D demonstrates that the quantitative accuracy of the initial MCR-ALS model decreases as R_s decreases, *S/N* decreases (dashed lines to solid lines), and target-interferent mass spectral similarity increases. For the *S/N* of 50 simulations, the percent error in modeled peak area reaches a maximum between 22 % (for low similarity pairs) and 37 % (for high similarity pairs). Notably, at a *S/N* of 10, the quantitative accuracy for the target analyte drastically decreases for a $R_s < 0.2$, where the percent error in peak area tops out between 77 % (for the low similarity pairs) and 105 % (for the high similarity pairs). The percent error for the initial MCR-

ALS model under these scenarios is due to chemometric multiplex disadvantage, causing those *m/z* signals from the interferent and noise to be wrongly attributed to the target analyte. Ultimately, the results in Figure 9.9C-D can be used to determine the R_s limit for achieving accurate identification (MV ≥ 800) and satisfactory quantitation results with the initial MCR-ALS model. Based on the *S/N* 10 simulations, which represents the limit of quantitation, the performance of the initial MCR-ALS models started to deteriorate below a R_s of 0.25. This R_s limit for MCR-ALS is consistent with previous work [19,21].

Conversely, Figure 9.9E-F illustrates (E) the average MV and (F) the average percent error in peak area (Eq. 9.2) of the target analyte using mzCompare *assisted* MCR-ALS. For all R_s and *S/N* levels simulated in this study, the average MV obtained with mzCompare assisted MCR-ALS is greater than 800 (Figure 9.9E). The improvement in the MV obtained for the target analytes under these different chromatographic conditions is because mzCompare can generate the pure elution profiles for both the target and interferent analytes by discovering their selective *m/z a priori*. Use of these pure elution profiles as constraints in the MCR-ALS model then allows the model to determine which *m/z* signals fit which analyte profile, suppressing the chemometric multiplex disadvantage in identification efforts. Overcoming the chemometric multiplex disadvantage with mzCompare assisted MCR-ALS also translates into the quantitative efforts (Figure 9.9F). Figure 9.9F shows that the maximum percent error in the target peak area ranges from 20 % to 37 % depending on the mass spectral similarity between the analytes and *S/N* simulated. Based on the results shown in Figure 9.9E-F and our experimental findings, the R_s limit of mzCompare assisted MCR-ALS is ~ 0.02. Ultimately, this work highlights the unique capabilities of mzCompare to enhance analyte discovery efforts for a single chromatogram due to its enhancement in R_s and n_c.

9.4. Conclusion

The mzCompare algorithm presented herein is a powerful computational tool to improve the identification and quantitation of overlapped analytes with MCR-ALS. This algorithm performs an intra-chromatogram comparison of the t_R and peak shape measured on each *m/z* to discover pure, selective *m/z* for each analyte in a sample. The elution profile of the selective *m/z* discovered by mzCompare can then improve MCR-ALS modeling when incorporated as an equality constraint. Using experimentally collected chromatograms, mzCompare assisted MCR-ALS could confidently identify overlapped analytes (MV $\geq$ 800) at a R_s as low as 0.05. The performance of mzCompare assisted MCR-ALS was thoroughly tested using GC-MS simulations at different R_s, *S/N* values, and degrees of mass spectra similarity between the target and interferent analytes. Confident analyte identification and satisfactory quantitative error for mzCompare assisted MCR-ALS was observed at a R_s as low as 0.02. In comparison, both the experimental and simulated data shows that the initial MCR-ALS models developed without this equality constraint struggled for regions affected by with severe chromatographic overlap (R_s < 0.3) and a high degree of spectrum contamination from noise/background interferences. This result is because these initial MCR-ALS models are affected by both rotational ambiguity and the chemometric multiplex disadvantage. Fortunately, mzCompare can overcome these modeling challenges to improve analyte identification and quantitation without requiring prior knowledge about the analytes of interest or the collection of multiple chromatograms. Future extensions of this work will explore the application of mzCompare to comprehensive two-dimensional gas chromatography-time-of-flight mass spectrometry (GC×GC-TOFMS) data and other chemometric decomposition approaches.

9.5. References

[1] Z. Wang, M. Fingas, L. Sigouin, Using multiple criteria for fingerprinting unknown oil samples having very similar chemical composition, Environ. Forensics. 3 (2002) 251–262. https://doi.org/10.1006/enfo.2002.0098.

[2] C.C. Chua, P. Brunswick, H. Kwok, J. Yan, D. Cuthbertson, G. Van Aggelen, D. Shang, Tiered approach to long-term weathered lubricating oil analysis: GC/FID, GC/MS diagnostic ratios, and multivariate statistics, Anal. Methods. 12 (2020) 5236–5246. https://doi.org/10.1039/d0ay01510e.

[3] P.E. Sudol, K.M. Pierce, S.E. Prebihalo, K.J. Skogerboe, B.W. Wright, R.E. Synovec, Development of gas chromatographic pattern recognition and classification tools for compliance and forensic analyses of fuels: A review, Anal. Chim. Acta. 1132 (2020) 157–186. https://doi.org/10.1016/j.aca.2020.07.027.

[4] C.W. Taylor, S.A. Bowden, What about nitrogen? Using nitrogen as a carrier gas during the analysis of petroleum biomarkers by gas chromatography mass spectrometry, J. Chromatogr. A. 1697 (2023) 463989. https://doi.org/10.1016/j.chroma.2023.463989.

[5] S. Yang, J.C. Hoggard, M.E. Lidstrom, R.E. Synovec, Comprehensive discovery of 13C labeled metabolites in the bacterium Methylobacterium extorquens AM1 using gas chromatography-mass spectrometry, J. Chromatogr. A. 1317 (2013) 175–185. https://doi.org/10.1016/j.chroma.2013.08.059.

[6] B.J. Webb-Robertson, Y.M. Kim, E.M. Zink, K.A. Hallaian, Q. Zhang, R. Madupu, K.M. Waters, T.O. Metz, A statistical analysis of the effects of urease pre-treatment on the measurement of the urinary metabolome by gas chromatography-mass spectrometry, Metabolomics. 10 (2014) 897–908. https://doi.org/10.1007/s11306-014-0642-1.

[7] O. Fiehn, Metabolomics by gas chromatography-mass spectometry: the combination of targeted and untargeted profiling, Curr. Protoc. Mol. Biol. 114 (2016) 30.4.1-30.4.32. https://doi.org/10.1186/s40945-017-0033-9.Using.

[8] G. Moros, A.C. Chatziioannou, H.G. Gika, N. Raikos, G. Theodoridis, Investigation of the derivatization conditions for GC-MS metabolomics of biological samples, Bioanalysis. 9 (2017) 53–65. https://doi.org/10.4155/bio-2016-0224.

[9] A. Ziółkowska, E. Wąsowicz, H.H. Jeleń, Differentiation of wines according to grape variety and geographical origin based on volatiles profiling using SPME-MS and SPME-GC/MS methods, Food Chem. 213 (2016) 714–720. https://doi.org/10.1016/j.foodchem.2016.06.120.

[10] D. Bressanello, E. Liberto, C. Cordero, P. Rubiolo, G. Pellegrino, M.R. Ruosi, C. Bicchi, Coffee aroma: Chemometric comparison of the chemical information provided by three different samplings combined with GC–MS to describe the sensory properties in cup, Food Chem. 214 (2017) 218–226. https://doi.org/10.1016/j.foodchem.2016.07.088.

[11] H. Song, J. Liu, GC-O-MS technique and its applications in food flavor analysis, Food Res. Int. 114 (2018) 187–198. https://doi.org/10.1016/j.foodres.2018.07.037.

[12] C.N. Cain, N.J. Haughn, H.J. Purcell, L.C. Marney, R.E. Synovec, C.T. Thoumsin, S.C. Jackels, K.J. Skogerboe, Analytical Determination of the Severity of Potato Taste Defect

in Roasted East African Arabica Coffee, J. Agric. Food Chem. 69 (2021) 2253–2261. https://doi.org/10.1021/acs.jafc.1c00605.

[13] J.M. Davis, J.C. Giddings, Statistical theory of component overlap in multicomponent chromatograms, Anal. Chem. 55 (1983) 418–424. https://doi.org/10.1021/ac00254a003.

[14] F.A. Chiappini, M.R. Alcaraz, G.M. Escandar, H.C. Goicoechea, A.C. Olivieri, Chromatographic applications in the multi-way calibration field, Molecules. 26 (2021) 1–29. https://doi.org/10.3390/molecules26216357.

[15] R. Tauler, Multivariate curve resolution applied to second order data, Chemom. Intell. Lab. Syst. 30 (1995) 133–146. https://doi.org/10.1016/0169-7439(95)00047-X.

[16] C. Ruckebusch, L. Blanchet, Multivariate curve resolution: A review of advanced and tailored applications and challenges, Anal. Chim. Acta. 765 (2013) 28–36. https://doi.org/10.1016/j.aca.2012.12.028.

[17] A. de Juan, J. Jaumot, R. Tauler, Multivariate curve resolution (MCR): Solving the mixture analysis problem, Anal. Methods. 6 (2014) 4964–4976. https://doi.org/10.1039/C4AY00571F.

[18] A.C. Olivieri, A down-to-earth analyst view of rotational ambiguity in second-order calibration with multivariate curve resolution − a tutorial, Anal. Chim. Acta. 1156 (2021) 338206. https://doi.org/10.1016/j.aca.2021.338206.

[19] I.H.M. van Stokkum, K.M. Mullen, V. V. Mihaleva, Global analysis of multiple gas chromatography-mass spectrometry (GC/MS) data sets: A method for resolution of co-eluting components with comparison to MCR-ALS, Chemom. Intell. Lab. Syst. 95 (2009) 150–163. https://doi.org/10.1016/j.chemolab.2008.10.004.

[20] X. Domingo-Almenara, A. Perera, N. Ramírez, N. Cañellas, X. Correig, J. Brezmes, Compound identification in gas chromatography/mass spectrometry-based metabolomics by blind source separation, J. Chromatogr. A. 1409 (2015) 226–233. https://doi.org/10.1016/j.chroma.2015.07.044.

[21] D.K. Pinkerton, B.C. Reaser, K.L. Berrier, R.E. Synovec, Determining the probability of achieving a successful quantitative analysis for gas chromatography-mass spectrometry, Anal. Chem. 89 (2017) 9926–9933. https://doi.org/10.1021/acs.analchem.7b02230.

[22] G.S. Ochoa, P.E. Sudol, T.J. Trinklein, R.E. Synovec, Class comparison enabled mass spectrum purification for comprehensive two-dimensional gas chromatography with time-of-flight mass spectrometry, Talanta. 236 (2022) 122844. https://doi.org/10.1016/j.talanta.2021.122844.

[23] C.N. Cain, T.J. Trinklein, G.S. Ochoa, R.E. Synovec, Tile-Based Pairwise Analysis of GC × GC-TOFMS Data to Facilitate Analyte Discovery and Mass Spectrum Purification, Anal. Chem. 94 (2022) 5658–5666. https://doi.org/10.1021/acs.analchem.2c00223.

[24] E. Voigtman, J.D. Winefordner, The Multiplex Disadvantage and Excess Low-Frequency Noise, Appl. Spectrosc. 41 (1987) 1182–1184. https://doi.org/10.1366/0003702874447509.

[25] R. Tauler, Calculation of maximum and minimum band boundaries of feasible solutions for species profiles obtained by multivariate curve resolution, J. Chemom. 15 (2001) 627–646. https://doi.org/10.1002/cem.654.

[26] S. Beyramysoltan, R. Rajkó, H. Abdollahi, Investigation of the equality constraint effect on the reduction of the rotational ambiguity in three-component system using a novel grid search method, Anal. Chim. Acta. 791 (2013) 25–35. https://doi.org/10.1016/j.aca.2013.06.043.

[27] G. Ahmadi, R. Tauler, H. Abdollahi, Multivariate calibration of first-order data with the correlation constrained MCR-ALS method, Chemom. Intell. Lab. Syst. 142 (2015) 143–150. https://doi.org/10.1016/j.chemolab.2014.11.010.

[28] B.D. Fitz, R.E. Synovec, Extension of the two-dimensional mass channel cluster plot method to fast separations utilizing low thermal mass gas chromatography with time-of-flight mass spectrometry, Anal. Chim. Acta. 913 (2016) 160–170. https://doi.org/10.1016/j.aca.2016.01.045.

[29] M. Ghaffari, A.C. Olivieri, H. Abdollahi, Strategy to Obtain Accurate Analytical Solutions in Second-Order Multivariate Calibration with Curve Resolution Methods, Anal. Chem. 90 (2018) 9725–9733. https://doi.org/10.1021/acs.analchem.8b00336.

[30] R.B. Pellegrino Vidal, F. Allegrini, A.C. Olivieri, The effect of constraints on the analytical figures of merit achieved by extended multivariate curve resolution-alternating least-squares, Anal. Chim. Acta. 1003 (2018) 10–15. https://doi.org/10.1016/j.aca.2017.12.008.

[31] W. Windig, J. Guilment, Interactive Self-Modeling Mixture Analysis, Anal. Chem. 63 (1991) 1425–1432.

[32] W. Windig, A. Bogomolov, S. Kucheryavskiy, Two-Way Data Analysis: Detection of Purest Variables, in: Compr. Chemom. Chem. Biochem. Data Anal., Second Edi, Elsevier, 2020: pp. 107–136. https://doi.org/10.1016/B978-0-12-409547-2.14747-X.

[33] C.N. Cain, S. Schöneich, R.E. Synovec, Development of an Enhanced Total Ion Current Chromatogram Algorithm to Improve Untargeted Peak Detection, Anal. Chem. 92 (2020) 11365–11373. https://doi.org/10.1021/acs.analchem.0c02136.

[34] Y. Izadmanesh, E. Garreta-Lara, J.B. Ghasemi, S. Lacorte, V. Matamoros, R. Tauler, Chemometric analysis of comprehensive two dimensional gas chromatography–mass spectrometry metabolomics data, J. Chromatogr. A. 1488 (2017) 113–125. https://doi.org/10.1016/j.chroma.2017.01.052.

[35] G.S. Ochoa, S.E. Prebihalo, B.C. Reaser, L.C. Marney, R.E. Synovec, Statistical inference of mass channel purity from Fisher ratio analysis using comprehensive two-dimensional gas chromatography with time of flight mass spectrometry data, J. Chromatogr. A. 1627 (2020) 461401. https://doi.org/10.1016/j.chroma.2020.461401.

[36] B.D. Fitz, B.C. Reaser, D.K. Pinkerton, J.C. Hoggard, K.J. Skogerboe, R.E. Synovec, Enhancing gas chromatography-time of flight mass spectrometry data analysis using two-dimensional mass channel cluster plots, Anal. Chem. 86 (2014) 3973–3979. https://doi.org/10.1021/ac5004344.

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

[38] J. Jaumot, R. Tauler, MCR-BANDS: A user friendly MATLAB program for the evaluation of rotation ambiguities in Multivariate Curve Resolution, Chemom. Intell. Lab. Syst. 103 (2010) 96–107. https://doi.org/10.1016/j.chemolab.2010.05.020.

[39] H.P. Bailey, S.C. Rutan, P.W. Carr, Factors that affect quantification of diode array data in comprehensive two-dimensional liquid chromatography using chemometric data analysis, J. Chromatogr. A. 1218 (2011) 8411–8422. https://doi.org/10.1016/j.chroma.2011.09.057.

[40] A. Eftekhari, H. Parastar, Multivariate analytical figures of merit as a metric for evaluation of quantitative measurements using comprehensive two-dimensional gas chromatography–mass spectrometry, J. Chromatogr. A. 1466 (2016) 155–165. https://doi.org/10.1016/j.chroma.2016.09.016.

[41] R.B. Pellegrino Vidal, A.C. Olivieri, R. Tauler, Quantifying the Prediction Error in Analytical Multivariate Curve Resolution Studies of Multicomponent Systems, Anal. Chem. 90 (2018) 7040–7047. https://doi.org/10.1021/acs.analchem.8b01431.

[42] X. Zhang, Z. Zhang, R. Tauler, Evaluation of the extension of rotation ambiguity associated to multivariate curve resolution solutions by the application of the MCR-BANDS method, Talanta. 202 (2019) 554–564. https://doi.org/10.1016/j.talanta.2019.05.002.

[43] B.D. Fitz, B.C. Mannion, K. To, T. Hoac, R.E. Synovec, Evaluation of injection methods for fast, high peak capacity separations with low thermal mass gas chromatography, J. Chromatogr. A. 1392 (2015) 82–90. https://doi.org/10.1016/j.chroma.2015.03.009.

Chapter 10: Conclusions and Future Directions

Both one-dimensional gas chromatography (1D-GC) and comprehensive two-dimensional gas chromatography (GC×GC) are versatile analytical techniques for resolving complex volatile and semi-volatile mixtures into their pure components. However, when these chromatographic platforms are coupled to mass spectrometry (MS), the resulting data is often too complex to manually analyze and interpret. Fortunately, the use of chemometrics can be used to extract meaningful chemical information from these chromatograms and ultimately, improve our understanding about an experimental study. The aim of this dissertation was to develop and apply novel chemometric algorithms that can discover and identify analytes of interest in 1D-GC-MS and GC×GC-MS chromatograms regardless of the analytical challenges faced. Based on this goal, various non-targeted chemometric methods were established herein for supervised, unsupervised, pairwise, and single chromatogram analyses.

10.1. Chapter 2 Summary and Future Directions

This chapter provided an initial characterization of the volatile fingerprint associated with potato taste defect (PTD) in roasted East African coffee beans. This flavor defect is caused by damage from the antestia bug, which feeds upon different parts of the coffee plant. Previous work had found that the vegetable-like odor that this defect is known for is due to the presence of 2-isopropyl-3-methoxypyrazine (IPMP). Hence, a primary focus of this chapter was to correlate the concentration of IPMP to the severity of PTD. Olfactory analysis of these coffee beans classified these beans as either having no off-odor (i.e., clean) or as having one of three different strength levels of PTD (mild, medium, and strong). The chemical profile of the headspace of these coffee beans was analyzed using one-dimensional gas chromatography-mass spectrometry

(1D-GC-MS). The concentration of IPMP was statistically different (p-value < 0.05) between the different odor categories, ranging from 0.6 – 3.1 ng/g in the clean samples, 1.6 – 72.4 ng/g in the mild PTD samples, 4.9 – 79.8 ng/g in the medium PTD samples, and 4.1 – 529.9 ng/g in the strong PTD samples. While these results demonstrated that the concentration of IPMP was positively correlated with severity of PTD, it is still unclear how damage from the antestia bug contributes to the concentration of IPMP in the coffee beans. Future work is needed to explore if IPMP formation in damaged coffee beans is from microorganisms carried by the pest or due to a stress response within the plant.

The second part of this chapter focused on the use of pixel-based Fisher ratio (F-ratio) analysis to determine if other volatiles in the headspace are affected by PTD. Twenty-one additional compounds with a statistical difference in concentration (p-value < 0.05) between the clean and strong PTD coffee samples were discovered. Volatiles with pleasant aromas in roasted coffee (e.g., fruity, nutty) were found in higher abundance in the clean samples while volatiles with negative aromas (e.g., musty, woody) were found in higher abundance in the strong PTD samples. This result was promising because it demonstrated that this flavor defect was affecting the entirety of the volatile headspace versus just a single compound. However, achieving a chemical fingerprint of PTD with 1D-GC-MS is challenging due to the limited peak capacity of these chromatograms. Therefore, these findings led to our future work, highlighted in Chapter 3, to reanalyze these samples with GC×GC-MS. Furthermore, collecting GC×GC-MS chromatograms of these samples would enable their analysis with the tile-based F-ratio algorithm. The tiling algorithm could mitigate the occurrence of false positives in the hit list due to retention time shifting and spurious instrumental noise, which fraught the pixel-based analysis of the 1D-GC-MS data in this chapter.

10.2. Chapter 3 Summary and Future Directions

As a follow-up to the previous chapter, Chapter 3 aims to thoroughly investigate the volatile headspace of these PTD-impacted coffee beans using GC×GC-MS and advanced chemometric methods. Tile-based F-ratio analysis discovered 327 analytes that had a statistically higher abundance in the clean coffee samples and 32 analytes, including IPMP, that had a statistically higher abundance in the strong PTD coffee samples (p-value < 0.01). The ~ 16-fold increase in the number of analytes discovered in this chapter is due to the increased resolving power of GC×GC-MS compared to 1D-GC-MS. Based on the analytes discovered, principal components analysis (PCA) was capable of distinguishing samples as clean or PTD-impacted. Also, a partial least squares (PLS) regression model to predict the concentration of IPMP using these statistically significant analytes was also developed. Some analytes that were negatively loaded in these chemometric models (i.e., had higher concentrations in the PTD-impacted samples) include 3-(2-cyclopentenyl)-2-methyl-1,1-diphenyl-1-propene, 2,4-diphenyl-4-methyl-2(*E*)-pentene, 2,4-di-*tert*-butylphenol, and 1,1,3-trimethyl-3-phenylindan. Concurrently, some analytes that were positively loaded in these chemometric models (e.g., had higher concentrations in the clean samples) were 2,6-dimethylpyrazine, 3-acetyl-2,5-dimethyl furan, furyl ethyl ketone, and 3-acetylpyrrole. The signal differences between the clean and strong PTD samples for these analytes could potentially be due to the presence of microorganisms on the coffee beans after antestia bug damage. However, as stated earlier, additional work is required to investigate this possible mechanism for PTD. It is also worth investigating the non-volatile portion of these coffee beans (both green and roasted) to further elucidate the biochemical mechanisms contributing to volatile fingerprint observed here.

10.3. Chapter 4 Summary and Future Directions

The aim for Chapter 4 was to discover how the chemical composition of a rocket fuel changes as it is exposed to the high temperatures associated with regenerative cooling engines. Using the Compact Rapid Assessment of Fuel Thermal Integrity (CRAFTI) apparatus, a highly paraffinic rocket fuel was exposed to four temperatures: 300, 500, 700, and 900 ^{o}F. The chemical composition of these fuels, along with their original sample prior to thermal stress exposure, was then characterized with GC×GC-MS and tile-based F-ratio analysis. In total, 92 analytes were found to have a statistically higher concentration (p-value < 0.01) in the fuels stressed at 900 ^{o}F compared to their original and 300 ^{o}F exposed counterparts. The compounds discovered in this work primarily belonged to compound classes known to form during the autooxidation and pyrolysis stages of fuel decomposition and cause the detrimental deposition of carbonaceous deposits inside engines (e.g., olefins, paraffins, aromatics, and oxygenated species). Even though the current literature suggests that these products are only formed at higher temperatures (> 662 ^{o}F), this chapter showed that several of these analytes were formed at temperatures as low at 300 ^{o}F. Hence, future work is required to understand the chemical reactions associated with the formation of these fuel decomposition products at lower thermal stress temperatures. It is also important to utilize the methodology in this chapter to understand how the chemical composition of other aerospace fuels change when exposed to thermal stress temperatures. Understanding these changes in chemical composition can ultimately provide more information into predicting fuel performance and potential engine failure from the buildup of carbonaceous deposits.

10.4. Chapter 5 Summary and Future Directions

Switching gears from supervised to unsupervised chemometric analyses, Chapter 5 introduces variance rank initiated-unsupervised sample indexing (VRI-USI). This approach

replaces the standard F-ratio metric calculated in peak table-, pixel-, or tile-based analyses with a metric of relative signal variance. The relative signal variance for a given analyte is due to two sources in an experiment, the background (e.g., natural variation, sample preparation, injection, and detection) and the chemical differences in a data set. Thus, if these background variance sources are known and/or controlled for, then a large signal variance indicates that there is some chemically relevant difference between samples. For analytes discovered to have a higher signal variance, *k*-means clustering is then performed to determine which samples are closely related. If the same samples are repeatedly clustered together as the analyst moves down the hit list, then this result would strongly suggest that the hidden relationship between the samples has been discovered in an unsupervised fashion. Application of VRI-USI to both simulated and experimentally collected metabolomics data sets showed that this approach has a similar performance to F-ratio analysis in discovering chemically meaningful differences between samples. Ultimately, this work enabled the capabilities of feature discovery to unsupervised analyses, where information regarding sample class membership is unknown. The next chapter in this dissertation (Chapter 6) provides the obvious next step of this work, where these principles are applied to a complex GC×GC-MS data set. However, another important future direction of this work is to further define these background sources of variation in an experiment. While natural variation between samples will always be present in an experiment, it is important to continue developing new sample preparation methods, injectors, and detectors that introduce minimal signal variance in comparative analyses.

10.5. Chapter 6 Summary and Future Directions

In Chapter 6, tile-based variance ranking is utilized as an unsupervised data reduction technique for GC×GC-MS in order to improve the modeling of three aerospace fuel properties

(viscosity, hydrogen content, and heat of combustion) with PLS regression. A data reduction strategy is commonly used prior to PLS modeling to remove noise and/or irrelevant signals from the data set, lowering prediction errors and improving computational speed. For instance, a common strategy implemented is to bin GC×GC-MS chromatograms down using a single-grid scheme. This chapter compares the accuracy of the PLS models developed with either single-grid binning or tile-based variance ranking using two metrics: the normalized root mean square error of cross-validation (NRMSECV) and prediction (NRMSEP). PLS models of viscosity, hydrogen content, and heat of combustion developed after binning the chromatograms had a NRMSECV between 12.1 – 14.4 % and a NRMSEP between 11.0 – 14.3 %. However, the models developed using the 521 analytes discovered by tile-based variance ranking were slightly more accurate, with a NRMSECV between 8.3 – 13.1 % and a NRMSEP between 7.6 – 13.5 %. To further improve upon the performance of these PLS models, a machine learning algorithm known as RReliefF was employed to discover which of the 521 analytes were best correlated with viscosity, hydrogen content, and heat of combustion. The PLS models after this data optimization strategy were highly accurate, with a NRMSECV and NRMSEP less than 8.4 %. Along with these highly accurate PLS models, another benefit of the tile-based variance ranking platform for data reduction is that the analyst can easily identify the analytes that are highly loaded in the PLS model. This advantage can facilitate a deeper understanding of the relationship between chemical composition and a given property measurement. Future work can continue to extend the methodology presented in Chapter 6 to property-composition modeling in other applications or to different chemometric models. Another potential study could be to compare the performance of PLS modeling using features discovered with tile-based variance ranking like in this chapter to those discovered with a supervised technique like tile-based F-ratio analysis

(see Chapter 3). Lastly, while the RReliefF feature optimization strategy employed herein was successful in decreasing the number of analytes down to those that were strongly correlated with each fuel property. It may also be worth exploring other machine learning algorithms to determine if another method can ultimately provide even lower PLS prediction errors.

10.6. Chapter 7 Summary and Future Directions

Chapter 7 presents the development of a new pairwise analysis method for GC×GC-MS chromatograms, referred to as tile-based 1v1 analysis. This work represents another transition point in this dissertation, where now the goal is to try to extract out the most chemical information about a sample using as few chromatograms as possible. Tile-based 1v1 analysis calculates the sum-normalized difference between two chromatograms to discover potential compounds of interest. Using a diesel fuel spiked with 18 non-native compounds, the performance of tile-based 1v1 analysis was similar to tile-based F-ratio analysis and superior to other pixel-based pairwise analysis approaches. A similar performance between tile-based 1v1 analysis and tile-based F-ratio analysis was also observed in the analysis of GC×GC-MS data set collected on cacao beans affected by moisture damage. Hence, this method can continue to be applied to other resource-limited or scouting studies.

Furthermore, this work demonstrated that tile-based 1v1 analysis could be readily coupled to class comparison enabled-mass spectrum purification (CCE-MSP) to improve analyte identification. CCE-MSP extracts the pure mass spectrum for the target analyte by subtracting the spectra from two chromatograms after their signals have been normalized to an interferent mass channel (*m/z*). Standard chemometric decomposition methods like multivariate curve resolution-alternating least squares (MCR-ALS) and parallel factor analysis (PARAFAC) were incapable of providing confident analyte identifications for 8 of the 18 spiked analytes. The poor

performance of MCR-ALS and PARAFAC is due to the chemometric multiplex disadvantage, where the signal from larger, overlapped interferences are included in the decomposed mass spectrum for the target species to minimize the residual error in the model. However, the pure mass spectra obtained with CCE-MSP enabled the confident identification of 17 out of 18 spiked analytes, demonstrating that this method overcomes the chemometric multiplex disadvantage. These findings open a multitude of research directions regarding CCE-MSP and the chemometric multiplex disadvantage. For example, it would be pertinent to understand the chromatographic limitations of CCE-MSP in obtaining the pure analyte mass spectrum. Another potential research question, which Chapter 9 attempts to answer, is how to overcome the chemometric multiplex disadvantage with only a single chromatogram.

10.7. Chapter 8 Summary and Future Directions

Chapter 8 details the development and application of a non-targeted peak detection algorithm for GC×GC-MS chromatograms. In a standard total ion current (TIC) chromatogram, peaks that have a signal greater than the noise summed along the *m/z* are easily detected. However, a significant portion of peaks with signals below the limit of detection can go undetected in the standard TIC chromatogram. The enhanced TIC algorithm was developed herein to improve the detection of these peaks. This algorithm discovers regions of analytical signal on every *m/z* in the chromatogram and zeroes out the remaining background noise. The resulting data can then be summed along the *m/z* dimension to generate the enhanced TIC chromatogram. Using the separation of a 90-component test mixture at 1 ppm and 10 ppm, the reported algorithm had respective recovery rates of 62 % and 93 %. Additionally, the application of the enhanced TIC algorithm to the separation of a yeast extract enabled the detection of 33 – 64 % more peaks. A longstanding question in analytical chemistry revolves around the true

distribution of analyte concentrations in a complex sample. By discovering these previously "undetectable" peaks, this work found that an exponential distribution had the best fit to the signal response curves for the test mixtures and yeast cell extracts. This finding has enabled the development of chromatographic simulation approaches that more closely match real world samples. For instance, an exponential distribution of analyte signals was employed herein to simulate GC×GC-MS chromatograms at different noise levels. These simulations were used to test the performance of the enhanced TIC algorithm in the context of statistical overlap theory (SOT). In these simulations, the number of peaks detected after application of the enhanced TIC algorithm was consistent with the predictions made by SOT. Future work with the enhanced TIC algorithm should test its performance on liquid chromatography-mass spectrometry (LC-MS) data. LC-MS data presents an interesting challenge to peak detection because a high degree of background noise is introduced with the use of mobile phase additives and ion suppression can reduce the MS response for analytes of interest.

10.8. Chapter 9 Summary and Future Directions

In Chapter 9, a new computational algorithm, referred to as mzCompare, is developed to discover analytes at low chromatographic resolutions (R_s) in a single 1D-GC-MS chromatogram. The mzCompare method assumes that *m/z* belonging to the same analyte should have similar retention times and peak shapes. Based on this assumption, the mzCompare algorithm performs an intra-chromatogram lack-of-fit (*LOF*) calculation between *m/z* to discover pure analytes in overlapped regions of the chromatogram. The mzCompare algorithm can also successively be coupled with chemometric decomposition techniques like MCR-ALS to improve identification and quantitation. Using the selective *m/z* discovered by mzCompare, this method generates a pure elution profile for each analyte, which can be utilized as an equality constraint for

decomposition models. The application of mzCompare was demonstrated herein using both experimental and simulated 1D-GC-MS chromatograms. For experimentally collected separations of known test mixtures, mzCompare not only discovered every single analyte in the sample but also improved the identification of unresolved species at a R_s as low as 0.05. Furthermore, the mzCompare algorithm provided a 44 % increase in the number of analytes detected in a complex aerospace fuel sample. Chromatographic simulations of target-interferent analyte pairs also demonstrated the benefit of coupling mzCompare with MCR-ALS to improve identification and quantitation. The performance for MCR-ALS models without the equality constraint developed by mzCompare started to decline at R_s below 0.25. In contrast, mzCompare assisted MCR-ALS continued to provide excellent identification and quantitation results at a R_s as low as 0.02. Overall, the results provided in this chapter highlight how the mzCompare methodology can help an analyst overcome both rotational ambiguities and the chemometric multiplex disadvantage.

Future directions for this work involve extending the mzCompare algorithm for GC×GC-MS chromatograms and other chemometric methods. For instance, decomposition models like PARAFAC can only be performed on three-way data that is sufficiently trilinear. Because of its trilinearity requirement, each PARAFAC model is unique and unaffected by any modeling ambiguities. However, the chemometric multiplex disadvantage is still a major cause for the poor performance of PARAFAC on extracting pure instrumental responses for analytes with low R_s and/or high degree of spectrum contamination from noise/background interferences (see Chapter 7). Therefore, developing the pure GC×GC elution profile for analytes in an unresolved region of the chromatogram with the mzCompare method could also improve the performance of PARAFAC. There are two potential ways of applying the mzCompare methodology to GC×GC-

MS data. The first way would involve treating every modulation in the GC×GC-MS chromatogram separately (i.e., effectively as a 1D-GC-MS chromatogram) and combining the results from each modulation together later. The main disadvantage of this approach is that it would be computationally time expensive. On the other hand, the second way would involve applying a tile-based peak finder to the GC×GC-MS chromatogram and then applying mzCompare to the unfolded data within each of those tiles. While this approach would take less computational time, it is possible that retention time shifting from one modulation to the next could inflate *LOF* measurements and ultimately, affect analyte discovery.

Bibliography

Abrahamsson, V., Ristic, N., Franz, K., & Van Geem, K. (2017). Comprehensive two-dimensional gas chromatography in combination with pixel-based analysis for fouling tendency prediction. Journal of Chromatography A, 1501, 89–98. https://doi.org/10.1016/j.chroma.2017.04.021

Adutwum, L. A., & Harynuk, J. J. (2014). Unique Ion Filter: A Data Reduction Tool for GC/MS Data Preprocessing Prior to Chemometric Analysis. Analytical Chemistry, 86(15), 7726–7733. https://doi.org/10.1021/ac501660a

Ahmadi, G., Tauler, R., & Abdollahi, H. (2015). Multivariate calibration of first-order data with the correlation constrained MCR-ALS method. Chemometrics and Intelligent Laboratory Systems, 142, 143–150. https://doi.org/10.1016/j.chemolab.2014.11.010

Ahmed, A. G., Murungi, L. K., & Babin, R. (2016). Developmental biology and demographic parameters of antestia bug Antestiopsis thunbergii (Hemiptera: Pentatomidae), on Coffea arabica (Rubiaceae) at different constant temperatures. International Journal of Tropical Insect Science, 36(3), 119–127. https://doi.org/10.1017/S1742758416000072

Alborzi, E., Gadsby, P., Ismail, M. S., Sheikhansari, A., Dwyer, M. R., Meijer, A. J. H. M., et al. (2019). Comparative Study of the Effect of Fuel Deoxygenation and Polar Species Removal on Jet Fuel Surface Deposition. Energy & Fuels, 33(3), 1825–1836. https://doi.org/10.1021/acs.energyfuels.8b03468

Alexandrino, G. L., Malmborg, J., Augusto, F., & Christensen, J. H. (2019). Investigating weathering in light diesel oils using comprehensive two-dimensional gas chromatography–High resolution mass spectrometry and pixel-based analysis: Possibilities and limitations. Journal of Chromatography A, 1591, 155–161. https://doi.org/10.1016/j.chroma.2019.01.042

Aliakbarzadeh, G., Sereshti, H., & Parastar, H. (2016). Pattern recognition analysis of chromatographic fingerprints of Crocus sativus L. secondary metabolites towards source identification and quality control. Analytical and Bioanalytical Chemistry, 408(12), 3295–3307. https://doi.org/10.1007/s00216-016-9400-8

Allen, R. C., & Rutan, S. C. (2011). Investigation of interpolation techniques for the reconstruction of the first dimension of comprehensive two-dimensional liquid chromatography-diode array detector data. Analytica Chimica Acta, 705(1–2), 253–260. https://doi.org/10.1016/j.aca.2011.06.022

Allen, R. C., & Rutan, S. C. (2012). Semi-automated alignment and quantification of peaks using parallel factor analysis for comprehensive two-dimensional liquid chromatography-diode array detector data sets. Analytica Chimica Acta, 723, 7–17. https://doi.org/10.1016/j.aca.2012.02.019

de Almeida, V. E., de Sousa Fernandes, D. D., Diniz, P. H. G. D., de Araújo Gomes, A., Véras, G., Galvão, R. K. H., & Araujo, M. C. U. (2021). Scores selection via Fisher's discriminant power in PCA-LDA to improve the classification of food data. Food Chemistry, 363(June). https://doi.org/10.1016/j.foodchem.2021.130296

Almstetter, M. F., Oefner, P. J., & Dettmer, K. (2012). Comprehensive two-dimensional gas chromatography in metabolomics. Analytical and Bioanalytical Chemistry, 402(6), 1993–2013. https://doi.org/10.1007/s00216-011-5630-y

Altin, O., & Eser, S. (2001). Analysis of Carboneceous Deposits from Thermal Stressing of a JP-8 Fuel on Superalloy Foils in a Flow Reactor. Industrial & Engineering Chemistry Research, 40(2), 589–595. https://doi.org/10.1021/ie0004489

Amador-Muñoz, O., & Marriott, P. J. (2008). Quantification in comprehensive two-dimensional gas chromatography and a model of quantification based on selected summed modulated peaks. Journal of Chromatography A, 1184(1–2), 323–340. https://doi.org/10.1016/j.chroma.2007.10.041

Amer, M. W., Mitrevski, B., Roy Jackson, W., Chaffee, A. L., & Marriott, P. J. (2014). Multidimensional and comprehensive two-dimensional gas chromatography of dichloromethane soluble products from a high sulfur Jordanian oil shale. Talanta, 120, 55–63. https://doi.org/10.1016/j.talanta.2013.11.069

Amigo, J. M., Skov, T., Bro, R., Coello, J., & Maspoch, S. (2008). Solving GC-MS problems with PARAFAC2. TrAC - Trends in Analytical Chemistry, 27(8), 714–725. https://doi.org/10.1016/j.trac.2008.05.011

Andersen, P. C., & Bruno, T. J. (2005). Thermal decomposition kinetics of RP-1 rocket propellant. Industrial and Engineering Chemistry Research, 44(6), 1670–1676. https://doi.org/10.1021/ie048958g

Arthur, D., & Vassilvitskii, S. (2007). k-means++: The advantages of careful seeding. In Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms (pp. 1027–1035). Philadelphia, PA: Society for Industrial and Applied Mathematics.

ASTM D445-19. Standard Test Method for Kinematic Viscosity of Transparent and Opaque Liquids (and Calculation of Dynamic Viscosity). (2019). West Conshohocken, PA. Retrieved from www.astm.org

ASTM D4809-18. Standard Test Method for Heat of Combustion of Liquid Hydrocarbon Fuels by Bomb Calorimeter (Precision Method). (2018). West Conshohocken, PA. Retrieved from www.astm.org

ASTM D7171-16. Standard Test Method for Hydrogen Content of Middle Distillate Petroleum Products by Low-Resolution Pulsed Nuclear Magnetic Resonance Spectroscopy. (2016). West Conshohocken, PA. Retrieved from www.astm.org

Bahaghighat, H. D., Freye, C. E., & Synovec, R. E. (2019). Recent advances in modulator technology for comprehensive two dimensional gas chromatography. TrAC - Trends in Analytical Chemistry, 113, 379–391. https://doi.org/10.1016/j.trac.2018.04.016

Bai, L., Smuts, J., Schenk, J., Cochran, J., & Schug, K. A. (2018). Comparison of GC-VUV, GC-FID, and comprehensive two-dimensional GC–MS for the characterization of weathered and unweathered diesel fuels. Fuel, 214, 521–527. https://doi.org/10.1016/j.fuel.2017.11.053

Bailey, H. P., & Rutan, S. C. (2011). Chemometric resolution and quantification of four-way data arising from comprehensive 2D-LC-DAD analysis of human urine. Chemometrics

and Intelligent Laboratory Systems, 106(1), 131–141. https://doi.org/10.1016/j.chemolab.2010.07.008

Bailey, H. P., Rutan, S. C., & Carr, P. W. (2011). Factors that affect quantification of diode array data in comprehensive two-dimensional liquid chromatography using chemometric data analysis. Journal of Chromatography A, 1218(46), 8411–8422. https://doi.org/10.1016/j.chroma.2011.09.057

Ballabio, D., & Consonni, V. (2013). Classification tools in chemistry. Part 1: Linear models. PLS-DA. Analytical Methods, 5(16), 3790–3798. https://doi.org/10.1039/c3ay40582f

Barcaru, A., & Vivó-Truyols, G. (2016). Use of Bayesian Statistics for Pairwise Comparison of Megavariate Data Sets: Extracting Meaningful Differences between GCxGC-MS Chromatograms Using Jensen–Shannon Divergence. Analytical Chemistry, 88(4), 2096–2104. https://doi.org/10.1021/acs.analchem.5b03506

Barwick, V. J. (1999). Sources of uncertainty in gas chromatography and high-performance liquid chromatography. Journal of Chromatography A, 849(1), 13–33. https://doi.org/10.1016/S0021-9673(99)00537-3

Bean, H. D., Hill, J. E., & Dimandja, J. M. D. (2015). Improving the quality of biomarker candidates in untargeted metabolomics via peak table-based alignment of comprehensive two-dimensional gas chromatography-mass spectrometry data. Journal of Chromatography A, 1394, 111–117. https://doi.org/10.1016/j.chroma.2015.03.001

Becker, R., Dohla, B., Nitz, S., & Vitzthum, O. G. (1987). Identification of the “Peasy” Off-Flavour Note in Central African Coffees. In 12th International Scientific Colloquium of Coffee, Montreaux, Switzerland, 29 June - 3 July 1987 (pp. 203–215). Paris, France: Association for Science and Information on Coffee (ASIC).

Beckstrom, A. C., Humston, E. M., Snyder, L. R., Synovec, R. E., & Juul, S. E. (2011). Application of comprehensive two-dimensional gas chromatography with time-of-flight mass spectrometry method to identify potential biomarkers of perinatal asphyxia in a non-human primate model. Journal of Chromatography A, 1218(14), 1899–1906. https://doi.org/10.1016/j.chroma.2011.01.086

Begue, N. J., Cramer, J. A., Von Bargen, C., Myers, K. M., Johnson, K. J., & Morris, R. E. (2011). Automated method for determining hydrocarbon distributions in mobility fuels. Energy & Fuels, 25(4), 1617–1623. https://doi.org/10.1021/ef101635a

Berrier, K. L., Freye, C. E., Billingsley, M. C., & Synovec, R. E. (2020). Predictive Modeling of Aerospace Fuel Properties Using Comprehensive Two-Dimensional Gas Chromatography with Time-Of-Flight Mass Spectrometry and Partial Least Squares Analysis. Energy & Fuels, 34(4), 4084–4094. https://doi.org/10.1021/acs.energyfuels.9b04108

Bertrand, B., Boulanger, R., Dussert, S., Ribeyre, F., Berthiot, L., Descroix, F., & Joët, T. (2012). Climatic factors directly impact the volatile organic compound fingerprint in green Arabica coffee bean as well as coffee beverage quality. Food Chemistry, 135(4), 2575–2583. https://doi.org/10.1016/j.foodchem.2012.06.060

Beyramysoltan, S., Rajkó, R., & Abdollahi, H. (2013). Investigation of the equality constraint effect on the reduction of the rotational ambiguity in three-component system using a

novel grid search method. Analytica Chimica Acta, 791, 25–35. https://doi.org/10.1016/j.aca.2013.06.043

Bieniek, A., & Moga, A. (2000). An efficient watershed algorithm based on connected components. Pattern Recognition, 33(6), 907–916. https://doi.org/10.1016/S0031-3203(99)00154-5

Bigirimana, J., Adams, C. G., Gatarayiha, C. M., Muhutu, J. C., & Gut, L. J. (2019). Occurrence of potato taste defect in coffee and its relations with management practices in Rwanda. Agriculture, Ecosystems and Environment, 269, 82–87. https://doi.org/10.1016/j.agee.2018.09.022

Bigirimana, J., Gerard, A., Mota-Sanchez, D., & Gut, L. J. (2018). Options for Managing Antestiopsis thunbergii (Hemiptera: Pentatomidae) and the Relationship of Bug Density to the Occurrence of Potato Taste Defect in Coffee. Florida Entomologist, 101(4), 580–586. https://doi.org/10.1653/024.101.0418

Blumberg, S., Frank, O., & Hofmann, T. (2010). Quantitative studies on the influence of the bean roasting parameters and hot water percolation on the concentrations of bitter compounds in coffee brew. Journal of Agricultural and Food Chemistry, 58(6), 3720–3728. https://doi.org/10.1021/jf9044606

Booksh, K. S., & Kowalski, B. R. (1994). Theory of Analytical Chemistry. Analytical Chemistry, 66(15), 782–791. https://doi.org/10.1021/ac00087a001

Boutou, S., & Chatonnet, P. (2007). Rapid headspace solid-phase microextraction/gas chromatographic/mass spectrometric assay for the quantitative determination of some of the main odorants causing off-flavours in wine. Journal of Chromatography A, 1141(1), 1–9. https://doi.org/10.1016/j.chroma.2006.11.106

Bouyjou, B., Decazy, B., & Fourny, G. (1999). Removing the “potato taste” from Burundian Arabica. Plantations, Recherche, Developpement, 6(2), 107–115.

Boyaci, E., Rodríguez-Lafuente, Á., Gorynski, K., Mirnaghi, F., Souza-Silva, É. A., Hein, D., & Pawliszyn, J. (2015). Sample preparation with solid phase microextraction and exhaustive extraction approaches: Comparison for challenging cases. Analytica Chimica Acta, 873, 14–30. https://doi.org/10.1016/j.aca.2014.12.051

Bressanello, D., Liberto, E., Cordero, C., Rubiolo, P., Pellegrino, G., Ruosi, M. R., & Bicchi, C. (2017). Coffee aroma: Chemometric comparison of the chemical information provided by three different samplings combined with GC–MS to describe the sensory properties in cup. Food Chemistry, 214, 218–226. https://doi.org/10.1016/j.foodchem.2016.07.088

Bressanello, D., Liberto, E., Cordero, C., Sgorbini, B., Rubiolo, P., Pellegrino, G., et al. (2018). Chemometric Modeling of Coffee Sensory Notes through Their Chemical Signatures: Potential and Limits in Defining an Analytical Tool for Quality Control. Journal of Agricultural and Food Chemistry, 66(27), 7096–7109. https://doi.org/10.1021/acs.jafc.8b01340

Bro, R. (1996). Multiway calibration: Multilinear PLS. Journal of Chemometrics, 10(1), 47–61. https://doi.org/10.1002/(SICI)1099-128X(199601)10:1<47::AID-CEM400>3.0.CO;2-C

Bro, R. (1997). PARAFAC. Tutorial and applications. Chemometrics and Intelligent Laboratory Systems, 38(2), 149–171. https://doi.org/10.1016/S0169-7439(97)00032-4

Brown, C. D., & Davis, H. T. (2006). Receiver operating characteristics curves and related decision measures: A tutorial. Chemometrics and Intelligent Laboratory Systems, 80(1), 24–38. https://doi.org/10.1016/j.chemolab.2005.05.004

Brownie, C., Boos, D. D., & Hughes-Oliver, J. (1990). Modifying the t and ANOVA F Tests When Treatment Is Expected to Increase Variability Relative to Controls. Biometrics, 46(1), 259–266.

Brownlee, J. (2020). Data Preparation for Machine Learning: Data Cleaning, Feature Selection, and Data Transforms in Python. Machine Learning Mastery.

Bruno, T. J., Huber, M. L., & Lemmon, E. W. (2009). Effect of RP-1 compositional variability on thermophysical properties. Energy & Fuels, 23(11), 5550–5555. https://doi.org/10.1021/ef900597q

Bruno, T. J., & Widegren, J. A. (2009a). Thermal decomposition kinetics of kerosene-based rocket propellants. 1. comparison of RP-1 and RP-2. Energy & Fuels, 23(11), 5517–5522. https://doi.org/10.1021/ef900576g

Bruno, T. J., & Widegren, J. A. (2009b). Thermal decomposition kinetics of kerosene-based rocket propellants. 2. RP-2 with three additives. Energy & Fuels, 23(11), 5523–5528. https://doi.org/10.1021/ef900577k

Bruno, T. J., & Windom, B. C. (2011). Method and apparatus for the thermal stress of complex fluids: Application to fuels. Energy & Fuels, 25(6), 2625–2632. https://doi.org/10.1021/ef2004738

Bueno, P. A., & Seeley, J. V. (2004). Flow-switching device for comprehensive two-dimensional gas chromatography. Journal of Chromatography A, 1027, 3–10. https://doi.org/10.1016/j.chroma.2003.10.033

Burger, B. V., Snyman, T., Burger, W. J. G., & Van Rooyen, W. F. (2003). Thermal modulator array for analyte modulation and comprehensive two-dimensional gas chromatography. Journal of Separation Science, 26, 123–128. https://doi.org/10.1002/jssc.200390002

Cai, D., Zhang, C., & He, X. (2010). Unsupervised feature selection for Multi-Cluster data. Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 333–342. https://doi.org/10.1145/1835804.1835848

Cain, C. N., Schöneich, S., & Synovec, R. E. (2020). Development of an Enhanced Total Ion Current Chromatogram Algorithm to Improve Untargeted Peak Detection. Analytical Chemistry, 92(16), 11365–11373. https://doi.org/10.1021/acs.analchem.0c02136

Cain, C. N., Haughn, N. J., Purcell, H. J., Marney, L. C., Synovec, R. E., Thoumsin, C. T., et al. (2021). Analytical Determination of the Severity of Potato Taste Defect in Roasted East African Arabica Coffee. Journal of Agricultural and Food Chemistry, 69(7), 2253–2261. https://doi.org/10.1021/acs.jafc.1c00605

Cain, C. N., Sudol, P. E., Berrier, K. L., & Synovec, R. E. (2021). Development of variance rank initiated-unsupervised sample indexing for gas chromatography-mass spectrometry analysis. Talanta, 233, 122495. https://doi.org/10.1016/j.talanta.2021.122495

Cain, C. N., Trinklein, T. J., Ochoa, G. S., & Synovec, R. E. (2022). Tile-Based Pairwise Analysis of GC × GC-TOFMS Data to Facilitate Analyte Discovery and Mass Spectrum Purification. Analytical Chemistry, 94(14), 5658–5666. https://doi.org/10.1021/acs.analchem.2c00223

Cain, C. N., Ochoa, G. S., & Synovec, R. E. (2023). Enhancing partial least squares modeling of comprehensive two-dimensional gas chromatography time-of-flight mass spectrometry data by tile-based variance ranking. Journal of Chromatography A, 1694, 463920. https://doi.org/10.1016/j.chroma.2023.463920

Capone, D. L., Ristic, R., Pardon, K. H., & Jeffery, D. W. (2015). Simple quantitative determination of potent thiols at ultratrace levels in wine by derivatization and high-performance liquid chromatography-tandem mass spectrometry (hplc-ms/ms) analysis. Analytical Chemistry, 87(2), 1226–1231. https://doi.org/10.1021/ac503883s

Caporaso, N., Whitworth, M. B., Cui, C., & Fisk, I. D. (2018). Variability of single bean coffee volatile compounds of Arabica and robusta roasted coffees analysed by SPME-GC-MS. Food Research International, 108, 628–640. https://doi.org/10.1016/j.foodres.2018.03.077

Chakravarthy, R., Acharya, C., Savalia, A., Naik, G. N., Das, A. K., Saravanan, C., et al. (2018). Property Prediction of Diesel Fuel Based on the Composition Analysis Data by two-Dimensional Gas Chromatography. Energy & Fuels, 32(3), 3760–3774. https://doi.org/10.1021/acs.energyfuels.7b03822

Charrad, M., Ghazzali, N., Boiteau, V., & Niknafs, A. (2014). Nbclust: An R package for determining the relevant number of clusters in a data set. Journal of Statistical Software, 61(6), 1–36. https://doi.org/10.18637/jss.v061.i06

Chatterjee, S., Major, G. H., Paull, B., Rodriguez, E. S., Kaykhaii, M., & Linford, M. R. (2018). Using pattern recognition entropy to select mass chromatograms to prepare total ion current chromatograms from raw liquid chromatography–mass spectrometry data. Journal of Chromatography A, 1558, 21–28. https://doi.org/10.1016/j.chroma.2018.04.042

Chauhan, A., Goyal, M. K., & Chauhan, P. (2014). GC-MS Technique and its Analytical Applications in Science and Technology. Journal of Analytical & Bioanalytical Techniques, 5(6). https://doi.org/10.4172/2155-9872.1000222

Chiappini, F. A., Alcaraz, M. R., Escandar, G. M., Goicoechea, H. C., & Olivieri, A. C. (2021). Chromatographic applications in the multi-way calibration field. Molecules, 26(21), 1–29. https://doi.org/10.3390/molecules26216357

Chin, S. T., Eyres, G. T., & Marriott, P. J. (2011). Identification of potent odourants in wine and brewed coffee using gas chromatography-olfactometry and comprehensive two-dimensional gas chromatography. Journal of Chromatography A, 1218(42), 7487–7498. https://doi.org/10.1016/j.chroma.2011.06.039

Chua, C. C., Brunswick, P., Kwok, H., Yan, J., Cuthbertson, D., Van Aggelen, G., & Shang, D. (2020). Tiered approach to long-term weathered lubricating oil analysis: GC/FID, GC/MS diagnostic ratios, and multivariate statistics. Analytical Methods, 12(43), 5236–5246. https://doi.org/10.1039/d0ay01510e

Cialiè Rosso, M., Liberto, E., Spigolon, N., Fontana, M., Somenzi, M., Bicchi, C., & Cordero, C. (2018). Evolution of potent odorants within the volatile metabolome of high-quality hazelnuts (Corylus avellana L.): evaluation by comprehensive two-dimensional gas chromatography coupled with mass spectrometry. Analytical and Bioanalytical Chemistry, 410(15), 3491–3506. https://doi.org/10.1007/s00216-017-0832-6

Cialiè Rosso, M., Stilo, F., Squara, S., Liberto, E., Mai, S., Mele, C., et al. (2020). Exploring extra dimensions to capture saliva metabolite fingerprints from metabolically healthy and unhealthy obese patients by comprehensive two-dimensional gas chromatography featuring Tandem Ionization mass spectrometry. Analytical and Bioanalytical Chemistry. https://doi.org/10.1007/s00216-020-03008-6

Commodo, M., Wong, O., Fabris, I., Groth, C. P. T., & Gülder, Ö. L. (2010). Spectroscopic study of aviation jet fuel thermal oxidative stability. Energy & Fuels, 24(12), 6437–6441. https://doi.org/10.1021/ef1012837

Consonni, V., Baccolo, G., Gosetti, F., Todeschini, R., & Ballabio, D. (2021). A MATLAB toolbox for multivariate regression coupled with variable selection. Chemometrics and Intelligent Laboratory Systems, 213, 104313. https://doi.org/10.1016/j.chemolab.2021.104313

Cook, D. W., & Rutan, S. C. (2014). Chemometrics for the analysis of chromatographic data in metabolomics investigations. Journal of Chemometrics, 28(9), 681–687. https://doi.org/10.1002/cem.2624

Cook, D. W., Rutan, S. C., Stoll, D. R., & Carr, P. W. (2015). Two dimensional assisted liquid chromatography - a chemometric approach to improve accuracy and precision of quantitation in liquid chromatography using 2D separation, dual detectors, and multivariate curve resolution. Analytica Chimica Acta, 859, 87–95. https://doi.org/10.1016/j.aca.2014.12.009

Cook, D. W., Burnham, M. L., Harmes, D. C., Stoll, D. R., & Rutan, S. C. (2017). Comparison of multivariate curve resolution strategies in quantitative LCxLC: Application to the quantification of furanocoumarins in apiaceous vegetables. Analytica Chimica Acta, 961(June 2016), 49–58. https://doi.org/10.1016/j.aca.2017.01.047

Cordero, C., Kiefl, J., Schieberle, P., Reichenbach, S. E., & Bicchi, C. (2015). Comprehensive two-dimensional gas chromatography and food sensory properties: Potential and challenges. Analytical and Bioanalytical Chemistry, 407(1), 169–191. https://doi.org/10.1007/s00216-014-8248-z

Cordero, C., Kiefl, J., Reichenbach, S. E., & Bicchi, C. (2019). Characterization of odorant patterns by comprehensive two-dimensional gas chromatography: A challenge in omic studies. TrAC - Trends in Analytical Chemistry, 113, 364–378. https://doi.org/10.1016/j.trac.2018.06.005

Cordero, C., Guglielmetti, A., Bicchi, C., Liberto, E., Baroux, L., Merle, P., et al. (2019). Comprehensive two-dimensional gas chromatography coupled with time of flight mass spectrometry featuring tandem ionization: Challenges and opportunities for accurate fingerprinting studies. Journal of Chromatography A, 1597, 132–141. https://doi.org/10.1016/j.chroma.2019.03.025

Crucello, J., Miron, L. F. O., Ferreira, V. H. C., Nan, H., Marques, M. O. M., Ritschel, P. S., et al. (2018). Characterization of the aroma profile of novel Brazilian wines by solid-phase microextraction using polymeric ionic liquid sorbent coatings. Analytical and Bioanalytical Chemistry, 410(19), 4749–4762. https://doi.org/10.1007/s00216-018-1134-3

Cuesta Sánchez, F., Toft, J., Van den Bogaert, B., & Massart, D. L. (1996). Orthogonal projection approach applied to peak purity assessment. Analytical Chemistry, 68(1), 79–85. https://doi.org/10.1021/ac950496g

Czerny, M., & Grosch, W. (2000). Potent Odorants of Raw Arabica Coffee. Their Changes during Roasting. Journal of Agricultural and Food Chemistry, 48(3), 868–872. https://doi.org/10.1021/jf990609n

Davies, D. L., & Bouldin, D. W. (1979). A Cluster Separation Measure. IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI-1(2), 224–227. https://doi.org/10.1109/TPAMI.1979.4766909

Davis, J. M. (1991). Statistical theory of spot overlap in two-dimensional separations. Analytical Chemistry, 63(19), 2141–2152. https://doi.org/10.1021/ac00019a014

Davis, J. M. (1993). Statistical theory of spot overlap for n-dimensional separations. Analytical Chemistry, 65(15), 2014–2023. https://doi.org/10.1021/ac00063a015

Davis, J. M. (2005). Statistical-overlap theory for elliptical zones of high aspect ratio in comprehensive two-dimensional separations. Journal of Separation Science, 28(4), 347–359. https://doi.org/10.1002/jssc.200401798

Davis, J. M. (2012). New theory for distribution of minimum resolution in multi-component separations with noise/detection limits. Journal of Chromatography A, 1251, 1–9. https://doi.org/10.1016/j.chroma.2012.06.034

Davis, J. M., & Carr, P. W. (2009). Effective saturation: A more informative metric for comparing peak separation in one- and two-dimensional separations. Analytical Chemistry, 81(3), 1198–1207. https://doi.org/10.1021/ac801728k

Davis, J. M., & Giddings, J. C. (1983). Statistical theory of component overlap in multicomponent chromatograms. Analytical Chemistry, 55(3), 418–424. https://doi.org/10.1021/ac00254a003

Davis, J. M., & Giddings, J. C. (1984). Origin and characterization of deparatures from the statistical model of component-peak overlap in chromatography. Journal of Chromatography, 289, 277–298.

Davis, J. M., & Giddings, J. C. (1985). Statistical method for estimation of number of components from single complex chromatograms: Theory, computer-based testing, and analysis of errors. Analytical Chemistry, 57(12), 2168–2177. https://doi.org/10.1021/ac00289a002

Davis, J. M., Stoll, D. R., & Carr, P. W. (2008a). Dependence of effective peak capacity in comprehensive two-dimensional separations on the distribution of peak capacity between the two dimensions. Analytical Chemistry, 80(21), 8122–8134. https://doi.org/10.1021/ac800933z

Davis, J. M., Stoll, D. R., & Carr, P. W. (2008b). Effect of first-dimension undersampling on effective peak capacity in comprehensive two-dimensional separations. Analytical Chemistry, 80(2), 461–473. https://doi.org/10.1021/ac071504j

Davis, T. J., Firzli, T. R., Higgins Keppler, E. A., Richardson, M., & Bean, H. D. (2022). Addressing Missing Data in GC × GC Metabolomics: Identifying Missingness Type and Evaluating the Impact of Imputation Methods on Experimental Replication. Analytical Chemistry, 94(31), 10912–10920. https://doi.org/10.1021/acs.analchem.1c04093

Dewitt, M. J., Edwards, T., Shafer, L., Brooks, D., Striebich, R., Bagley, S. P., & Wornat, M. J. (2011). Effect of aviation fuel type on pyrolytic reactivity and deposition propensity under supercritical conditions. Industrial and Engineering Chemistry Research, 50(18), 10434–10451. https://doi.org/10.1021/ie200257b

Ding, C., & He, X. (2004). K-means clustering via principal component analysis. Proceedings, Twenty-First International Conference on Machine Learning, ICML 2004, 225–232. https://doi.org/10.1145/1015330.1015408

Domingo-Almenara, X., Perera, A., Ramírez, N., Cañellas, N., Correig, X., & Brezmes, J. (2015). Compound identification in gas chromatography/mass spectrometry-based metabolomics by blind source separation. Journal of Chromatography A, 1409, 226–233. https://doi.org/10.1016/j.chroma.2015.07.044

Dondi, F., Du Ale Kahie, Y., Lodi, G., Remelli, M., Reschiglian, P., & Bighi, C. (1986). Evaluation of the number of components in multi-component liquid chromatograms of plant extracts. Analytica Chimica Acta, 191, 261–273. https://doi.org/10.1016/S0003-2670(00)86313-8

Dondi, F., Bassi, A., Cavazzini, A., & Pietrogrande, M. C. (1998). A Quantitative Theory of the Statistical Degree of Peak Overlapping in Chromatography. Analytical Chemistry, 70(4), 766–773. https://doi.org/10.1021/ac9705430

Dubois, L. M., Perrault, K. A., Stefanuto, P.-H., Koschinski, S., Edwards, M., McGregor, L., & Focant, J.-F. (2017). Thermal desorption comprehensive two-dimensional gas chromatography coupled to variable-energy electron ionization time-of-flight mass spectrometry for monitoring subtle changes in volatile organic compound profiles of human blood. Journal of Chromatography A, 1501, 117–127. https://doi.org/10.1016/j.chroma.2017.04.026

Ebadzadsahrai, G., Higgins Keppler, E. A., Soby, S. D., & Bean, H. D. (2020). Inhibition of Fungal Growth and Induction of a Novel Volatilome in Response to Chromobacterium vaccinii Volatile Organic Compounds. Frontiers in Microbiology, 11. https://doi.org/10.3389/fmicb.2020.01035

Eftekhari, A., & Parastar, H. (2016). Multivariate analytical figures of merit as a metric for evaluation of quantitative measurements using comprehensive two-dimensional gas chromatography–mass spectrometry. Journal of Chromatography A, 1466, 155–165. https://doi.org/10.1016/j.chroma.2016.09.016

Enke, C. G., & Nagels, L. J. (2011). Undetected components in natural mixtures: How many? What concentrations? Do they account for chemical noise? What is needed to detect them? Analytical Chemistry, 83(7), 2539–2546. https://doi.org/10.1021/ac102818a

Eser, S., Venkataraman, R., & Altin, O. (2006). Deposition of carbonaceous solids on different substrates from thermal stressing of JP-8 and jet A fuels. Industrial and Engineering Chemistry Research, 45(26), 8946–8955. https://doi.org/10.1021/ie060968p

Fan, Z., Rahimi, P., Alem, T., Eisenhawer, A., & Arboleda, P. (2011). Fouling characteristics of hydrocarbon streams containing olefins and conjugated olefins. Energy & Fuels, 25(3), 1182–1190. https://doi.org/10.1021/ef101661n

Fang, S., Liu, S., Song, J., Huang, Q., & Xiang, Z. (2021). Recognition of pathogens in food matrixes based on the untargeted in vivo microbial metabolite profiling via a novel SPME/GC × GC-QTOFMS approach. Food Research International, 142, 110213. https://doi.org/10.1016/j.foodres.2021.110213

Farah, A., Monteiro, M. C., Calado, V., Franca, A. S., & Trugo, L. C. (2006). Correlation between cup quality and chemical attributes of Brazilian coffee. Food Chemistry, 98(2), 373–380. https://doi.org/10.1016/j.foodchem.2005.07.032

Feng, Y., Su, G., Zhao, H., Cai, Y., Cui, C., Sun-Waterhouse, D., & Zhao, M. (2015). Characterisation of aroma profiles of commercial soy sauce by odour activity value and omission test. Food Chemistry, 167, 220–228. https://doi.org/10.1016/j.foodchem.2014.06.057

Fiehn, O. (2016). Metabolomics by gas chromatography-mass spectometry: the combination of targeted and untargeted profiling. Current Protocols in Molecular Biology, 114, 30.4.1-30.4.32. https://doi.org/10.1186/s40945-017-0033-9.Using

Field, J. A., Nickerson, G., James, D. D., & Heider, C. (1996). Determination of essential oils in hops by headspace solid-phase microextraction. Journal of Agricultural and Food Chemistry, 44(7), 1768–1772. https://doi.org/10.1021/jf950663d

Fitz, B. D., & Synovec, R. E. (2016). Extension of the two-dimensional mass channel cluster plot method to fast separations utilizing low thermal mass gas chromatography with time-of-flight mass spectrometry. Analytica Chimica Acta, 913, 160–170. https://doi.org/10.1016/j.aca.2016.01.045

Fitz, B. D., Reaser, B. C., Pinkerton, D. K., Hoggard, J. C., Skogerboe, K. J., & Synovec, R. E. (2014). Enhancing gas chromatography-time of flight mass spectrometry data analysis using two-dimensional mass channel cluster plots. Analytical Chemistry, 86(8), 3973–3979. https://doi.org/10.1021/ac5004344

Fitz, B. D., Mannion, B. C., To, K., Hoac, T., & Synovec, R. E. (2015). Evaluation of injection methods for fast, high peak capacity separations with low thermal mass gas chromatography. Journal of Chromatography A, 1392, 82–90. https://doi.org/10.1016/j.chroma.2015.03.009

Fortin, T. J., & Bruno, T. J. (2013). Assessment of the thermophysical properties of thermally stressed RP-1 and RP-2. Energy & Fuels, 27(5), 2506–2514. https://doi.org/10.1021/ef400193d

Fraga, C. G. (2003). Chemometric approach for the resolution and quantification of unresolved peaks in gas chromatography-selected-ion mass spectrometry data. Journal of Chromatography A, 1019, 31–42. https://doi.org/10.1016/S0021-9673(03)01329-3

Franca, A. S., Oliveira, L. S., Mendonça, J. C. F., & Silva, X. A. (2005). Physical and chemical attributes of defective crude and roasted coffee beans. Food Chemistry, 90, 89–94. https://doi.org/10.1016/j.foodchem.2004.03.028

Franchina, F. A., Maimone, M., Tranchida, P. Q., & Mondello, L. (2016). Flow modulation comprehensive two-dimensional gas chromatography-mass spectrometry using ≈ 4 mL min-1 gas flows. Journal of Chromatography A, 1441, 134–139. https://doi.org/10.1016/j.chroma.2016.02.041

Frankenfeld, J. W., & Taylor, W. F. (1980). Deposit Formation from Deoxygenated Hydrocarbons. 4. Studies in Pure Compound Systems. Industrial and Engineering Chemistry Product Research and Development, 19(1), 65–70. https://doi.org/10.1021/i360073a016

Fränti, P., & Sieranoja, S. (2018). K-means properties on six clustering benchmark datasets. Applied Intelligence, 48(12), 4743–4759. https://doi.org/10.1007/s10489-018-1238-7

Frato, K. E. (2019). Identification of Hydroxypyrazine O-Methyltransferase Genes in Coffea arabica: A Potential Source of Methoxypyrazines That Cause Potato Taste Defect. Journal of Agricultural and Food Chemistry, 67(1), 341–351. https://doi.org/10.1021/acs.jafc.8b04541

Freye, C. E., & Synovec, R. E. (2016). High temperature diaphragm valve-based comprehensive two-dimensional gas chromatography with time-of-flight mass spectrometry. Talanta, 161, 675–680. https://doi.org/10.1016/j.talanta.2016.09.002

Freye, C. E., Fitz, B. D., Billingsley, M. C., & Synovec, R. E. (2016). Partial least squares analysis of rocket propulsion fuel data using diaphragm valve-based comprehensive two-dimensional gas chromatography coupled with flame ionization detection. Talanta, 153, 203–210. https://doi.org/10.1016/j.talanta.2016.03.016

Freye, C. E., Moore, N. R., & Synovec, R. E. (2018). Enhancing the chemical selectivity in discovery-based analysis with tandem ionization time-of-flight mass spectrometry detection for comprehensive two-dimensional gas chromatography. Journal of Chromatography A, 1537, 99–108. https://doi.org/10.1016/j.chroma.2018.01.008

Freye, C. E., Bowden, P. R., Greenfield, M. T., & Tappan, B. C. (2020). Non-targeted discovery-based analysis for gas chromatography with mass spectrometry: A comparison of peak table, tile, and pixel-based Fisher ratio analysis. Talanta, 211, 120668. https://doi.org/10.1016/j.talanta.2019.120668

Frysinger, G. S., & Gaines, R. B. (2002). Forensic analysis of ignitable liquids in fire debris by comprehensive two-dimensional gas chromatography. Journal of Forensic Sciences, 47(3), 471–482. https://doi.org/10.1520/JFS15288J

Furbo, S., Hansen, A. B., Skov, T., & Christensen, J. H. (2014). Pixel-based analysis of comprehensive two-dimensional gas chromatograms (color plots) of petroleum: A tutorial. Analytical Chemistry, 86(15), 7160–7170. https://doi.org/10.1021/ac403650d

Gao, M., Hou, L., Zhang, X., & Zhang, D. (2019). Coke deposition inhibition for endothermic hydrocarbon fuels in a reforming catalyst-coated reactor. Energy & Fuels, 33(7), 6126–6133. https://doi.org/10.1021/acs.energyfuels.9b00878

Ghaffari, M., Olivieri, A. C., & Abdollahi, H. (2018). Strategy to Obtain Accurate Analytical Solutions in Second-Order Multivariate Calibration with Curve Resolution Methods. Analytical Chemistry, 90(16), 9725–9733. https://doi.org/10.1021/acs.analchem.8b00336

Ghosh, A., Bates, C. T., Seeley, S. K., & Seeley, J. V. (2013). High speed Deans switch for low duty cycle comprehensive two-dimensional gas chromatography. Journal of Chromatography A, 1291, 146–154. https://doi.org/10.1016/j.chroma.2013.04.003

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

Giocastro, B., Zoccali, M., Tranchida, P. Q., & Mondello, L. (2021). Evaluation of different internal diameter coated modulation columns within the context of solid-state modulation. Journal of Separation Science, 1–20. https://doi.org/10.1002/jssc.202001031

Di Giovanni, N., Meuwis, M. A., Louis, E., & Focant, J. F. (2020). Untargeted Serum Metabolic Profiling by Comprehensive Two-Dimensional Gas Chromatography-High-Resolution Time-of-Flight Mass Spectrometry. Journal of Proteome Research, 19(3), 1013–1028. https://doi.org/10.1021/acs.jproteome.9b00535

Gosetti, F., Mazzucco, E., Zampieri, D., & Gennaro, M. C. (2010). Signal suppression/enhancement in high-performance liquid chromatography tandem mass spectrometry. Journal of Chromatography A, 1217(25), 3929–3937. https://doi.org/10.1016/j.chroma.2009.11.060

Gough, R. V., & Bruno, T. J. (2013). Composition-explicit distillation curves of alternative turbine fuels. Energy & Fuels, 27(1), 294–302. https://doi.org/10.1021/ef3016848

Griffith, J. F., Winniford, W. L., Sun, K., Edam, R., & Luong, J. C. (2012). A reversed-flow differential flow modulator for comprehensive two-dimensional gas chromatography. Journal of Chromatography A, 1226, 116–123. https://doi.org/10.1016/j.chroma.2011.11.036

Gruber, B., Weggler, B. A., Jaramillo, R., Murrell, K. A., Piotrowski, P. K., & Dorman, F. L. (2018). Comprehensive two-dimensional gas chromatography in forensic science: A critical review of recent trends. TrAC - Trends in Analytical Chemistry, 105, 292–301. https://doi.org/10.1016/j.trac.2018.05.017

Gueule, D., Fourny, G., Ageron, E., Le Flèche-Matéos, A., Vandenbogaert, M., Grimont, P. A. D., & Cilas, C. (2015). Pantoea coffeiphila sp. nov., cause of the ‘potato taste’ of Arabica coffee from the African great lakes region. International Journal of Systematic and Evolutionary Microbiology, 65(1), 23–29. https://doi.org/10.1099/ijs.0.063545-0

Gundlach-Graham, A., & Enke, C. G. (2015). Effect of response factor variations on the response distribution of complex mixtures. European Journal of Mass Spectrometry, 21(3), 471–479. https://doi.org/10.1255/ejms.1369

Guozhu, L., Yongjin, H., Li, W., Xiangwen, Z., & Zhentao, M. (2009). Solid deposits from thermal stressing of n - Dodecane and chinese RP-3 jet fuel in the presence of several initiators. Energy & Fuels, 23(1), 356–365. https://doi.org/10.1021/ef800657z

Haar, L., Anding, K., Trambitckii, K., & Notni, G. (2019). Comparison between supervised and unsupervised feature selection methods. ICPRAM 2019 - Proceedings of the 8th International Conference on Pattern Recognition Applications and Methods, 582–589. https://doi.org/10.5220/0007385305820589

Haglund, P., Korytár, P., Danielsson, C., Diaz, J., Wiberg, K., Leonards, P., et al. (2008). GC×GC-ECD: A promising method for the determination of dioxins and dioxin-like PCBs in food and feed. Analytical and Bioanalytical Chemistry, 390(7), 1815–1827. https://doi.org/10.1007/s00216-008-1896-0

Hale, A. R., Ruegger, P. M., Rolshausen, P., Borneman, J., & Yang, J. in. (2022). Fungi associated with the potato taste defect in coffee beans from Rwanda. Botanical Studies, 63(1). https://doi.org/10.1186/s40529-022-00346-9

Hameed, A., Hussain, S. A., Ijaz, M. U., Ullah, S., Pasha, I., & Suleria, H. A. R. (2018). Farm to Consumer: Factors Affecting the Organoleptic Characteristics of Coffee. II: Postharvest Processing Factors. Comprehensive Reviews in Food Science and Food Safety, 17(5), 1184–1237. https://doi.org/10.1111/1541-4337.12365

Han, S., Zhang, W., Li, P., Li, X., Liu, J., Xu, B., & Luo, D. (2017). Characterization of Aromatic Liquor by Gas Chromatography and Principal Component Analysis. Analytical Letters, 50(5), 777–786. https://doi.org/10.1080/00032719.2016.1196365

Hantao, L. W., Toledo, B. R., De Lima Ribeiro, F. A., Pizetta, M., Pierozzi, C. G., Furtado, E. L., & Augusto, F. (2013). Comprehensive two-dimensional gas chromatography combined to multivariate data analysis for detection of disease-resistant clones of Eucalyptus. Talanta, 116, 1079–1084. https://doi.org/10.1016/j.talanta.2013.08.033

Harshman, R. A., & Lundy, M. E. (1994). PARAFAC: Parallel factor analysis. Computational Statistics and Data Analysis, 18(1), 39–72. https://doi.org/10.1016/0167-9473(94)90132-5

Harvey, P. M., & Shellie, R. A. (2011). Factors affecting peak shape in comprehensive two-dimensional gas chromatography with non-focusing modulation. Journal of Chromatography A, 1218(21), 3153–3158. https://doi.org/10.1016/j.chroma.2010.08.029

Haug, K., Salek, R. M., Conesa, P., Hastings, J., De Matos, P., Rijnbeek, M., et al. (2013). MetaboLights - An open-access general-purpose repository for metabolomics studies and associated meta-data. Nucleic Acids Research, 41(D1), 781–786. https://doi.org/10.1093/nar/gks1004

He, X., Cai, D., & Niyogi, P. (2006). Laplacian Score for Feature Selection. In Y. Weiss, B. Schölkopf, & J. C. Platt (Eds.), Advances in Neural Information Processing Systems 18 (pp. 507–514). MIT Press.

Heinemann, J. (2019). Machine Learning in Untargeted Metabolomics Experiments. In E. E. K. Baidoo (Ed.), Microbial Metabolomics: Methods and Protocols (pp. 287–299). New York, NY: Springer New York. https://doi.org/10.1007/978-1-4939-8757-3_17

Henrion, R. (1994). N-way principal component analysis theory, algorithms and applications. Chemometrics and Intelligent Laboratory Systems, 25(1), 1–23. https://doi.org/10.1016/0169-7439(93)E0086-J

Herman, D. P., Gonnord, M. F., & Guiochon, G. (1984). Statistical approach for estimating the total number of components in complex mixtures from nontotally resolved chromatograms. Analytical Chemistry, 56(6), 995–1003. https://doi.org/10.1021/ac00270a030

Heyne, J., Bell, D., Feldhausen, J., Yang, Z., & Boehm, R. (2022). Towards fuel composition and properties from Two-dimensional gas chromatography with flame ionization and vacuum ultraviolet spectroscopy. Fuel, 312, 122709. https://doi.org/10.1016/j.fuel.2021.122709

Hoggard, J. C., & Synovec, R. E. (2007). Parallel factor analysis (PARAFAC) of target analytes in GC × GC-TOFMS data: Automated selection of a model with an appropriate number of factors. Analytical Chemistry, 79(4), 1611–1619. https://doi.org/10.1021/ac061710b

Hoggard, J. C., Wahl, J. H., Synovec, R. E., Mong, G. M., & Fraga, C. G. (2010). Impurity Profiling of a Chemical Weapon Precursor for Possible Forensic Signatures by Comprehensive Two-Dimensional Gas Chromatography/Mass Spectrometry and Chemometrics. Analytical Chemistry, 82(2), 689–698. https://doi.org/10.1021/ac902247x

Hoh, E., Dodder, N. G., Lehotay, S. J., Pangallo, K. C., Reddy, C. M., & Maruya, K. A. (2012). Nontargeted comprehensive two-dimensional gas chromatography/time-of-flight mass spectrometry method and software for inventorying persistent and bioaccumulative contaminants in marine environments. Environmental Science and Technology, 46(15), 8001–8008. https://doi.org/10.1021/es301139q

Holscher, W., & Steinhart, H. (1995). Aroma Compounds in Green Coffee. Developments in Food Science, 37(C), 785–803. https://doi.org/10.1016/S0167-4501(06)80196-2

Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). Assessing the Fit of the Model. In Applied Logistic Regression (3rd ed., pp. 153–226). Hoboken, NJ: Wiley.

Hsieh, P. Y., Abel, K. R., & Bruno, T. J. (2013). Analysis of marine diesel fuel with the advanced distillation curve method. Energy & Fuels, 27(2), 804–810. https://doi.org/10.1021/ef3020525

Huang, L. F., Wu, M. J., Zhong, K. J., Sun, X. J., Liang, Y. Z., Dai, Y. H., et al. (2007). Fingerprint developing of coffee flavor by gas chromatography-mass spectrometry and combined chemometrics methods. Analytica Chimica Acta, 588(2), 216–223. https://doi.org/10.1016/j.aca.2007.02.013

Humston, E. M., Knowles, J. D., McShea, A., & Synovec, R. E. (2010). Quantitative assessment of moisture damage for cacao bean quality using two-dimensional gas chromatography combined with time-of-flight mass spectrometry and chemometrics. Journal of Chromatography A, 1217(12), 1963–1970. https://doi.org/10.1016/j.chroma.2010.01.069

Huzel, D. K., & Huang, D. H. (1992). Modern Engineering for Design of Liquid-Propellant Rocket Engines. Washington, DC: American Institute of Aeronautics and Astronautics, Inc.

Izadmanesh, Y., Garreta-Lara, E., Ghasemi, J. B., Lacorte, S., Matamoros, V., & Tauler, R. (2017). Chemometric analysis of comprehensive two dimensional gas chromatography–mass spectrometry metabolomics data. Journal of Chromatography A, 1488, 113–125. https://doi.org/10.1016/j.chroma.2017.01.052

Jackels, S. C., Marshall, E. E., Omaiye, A. G., Gianan, R. L., Lee, F. T., & Jackels, C. F. (2014). GCMS investigation of volatile compounds in green coffee affected by potato taste defect and the antestia bug. Journal of Agricultural and Food Chemistry, 62(42), 10222–10229. https://doi.org/10.1021/jf5034416

Jackson, D. A. (1993). Stopping Rules in Principal Components Analysis: A Comparison of Heuristical and Statistical Approaches. Ecology, 74(8), 2204–2214. https://doi.org/10.2307/1939574

Jain, A. K. (2010). Data clustering: 50 years beyond K-means. Pattern Recognition Letters, 31(8), 651–666. https://doi.org/10.1016/j.patrec.2009.09.011

Jaumot, J., & Tauler, R. (2010). MCR-BANDS: A user friendly MATLAB program for the evaluation of rotation ambiguities in Multivariate Curve Resolution. Chemometrics and Intelligent Laboratory Systems, 103(2), 96–107. https://doi.org/10.1016/j.chemolab.2010.05.020

Jennerwein, M. K., Eschner, M., Gröger, T., Wilharm, T., & Zimmermann, R. (2014). Complete group-type quantification of petroleum middle distillates based on comprehensive two-dimensional gas chromatography time-of-flight mass spectrometry (GCxGC-TOFMS) and visual basic scripting. Energy & Fuels, 28(9), 5670–5681. https://doi.org/10.1021/ef501247h

Jennerwein, M. K., Sutherland, A. C., Eschner, M., Gröger, T., Wilharm, T., & Zimmermann, R. (2017). Quantitative analysis of modern fuels derived from middle distillates – The impact of diverse compositions on standard methods evaluated by an offline hyphenation of HPLC-refractive index detection with GC×GC-TOFMS. Fuel, 187, 16–25. https://doi.org/10.1016/j.fuel.2016.09.033

Jennerwein, M. K., Eschner, M., Wilharm, T., Gröger, T., & Zimmermann, R. (2019). Evaluation of reversed phase versus normal phase column combination for the quantitative analysis of common commercial available middle distillates using GC × GC-TOFMS and Visual Basic Script. Fuel, 235, 336–338. https://doi.org/10.1016/j.fuel.2018.07.081

Jha, S. K., & Hayashi, K. (2017). Molecular structural discrimination of chemical compounds in body odor using their GC–MS chromatogram and clustering methods. International Journal of Mass Spectrometry, 423, 1–14. https://doi.org/10.1016/j.ijms.2017.09.010

Jin, B., Jing, K., Liu, J., Zhang, X., & Liu, G. (2017). Pyrolysis and coking of endothermic hydrocarbon fuel in regenerative cooling channel under different pressures. Journal of Analytical and Applied Pyrolysis, 125, 117–126. https://doi.org/10.1016/j.jaap.2017.04.010

Johnson, K. J., & Synovec, R. E. (2002). Pattern recognition of jet fuels: Comprehensive GC × GC with ANOVA-based feature selection and principal component analysis. Chemometrics and Intelligent Laboratory Systems, 60, 225–237. https://doi.org/10.1016/S0169-7439(01)00198-8

Johnson, K. J., Wright, B. W., Jarman, K. H., & Synovec, R. E. (2003). High-speed peak matching algorithm for retention time alignment of gas chromatographic data for

chemometric analysis. Journal of Chromatography A, 996(1–2), 141–155. https://doi.org/10.1016/S0021-9673(03)00616-2

de Juan, A., Jaumot, J., & Tauler, R. (2014). Multivariate curve resolution (MCR): Solving the mixture analysis problem. Analytical Methods, 6(14), 4964–4976. https://doi.org/10.1039/C4AY00571F

Kallio, M., Kivilompolo, M., Varjo, S., Jussila, M., & Hyötyläinen, T. (2009). Data analysis programs for comprehensive two-dimensional chromatography. Journal of Chromatography A, 1216(14), 2923–2927. https://doi.org/10.1016/j.chroma.2008.11.037

Kehimkar, B., Hoggard, J. C., Marney, L. C., Billingsley, M. C., Fraga, C. G., Bruno, T. J., & Synovec, R. E. (2014). Correlation of rocket propulsion fuel properties with chemical composition using comprehensive two-dimensional gas chromatography with time-of-flight mass spectrometry followed by partial least squares regression analysis. Journal of Chromatography A, 1327, 132–140. https://doi.org/10.1016/j.chroma.2013.12.060

Kehimkar, B., Parsons, B. A., Hoggard, J. C., Billingsley, M. C., Bruno, T. J., & Synovec, R. E. (2015). Modeling RP-1 fuel advanced distillation data using comprehensive two-dimensional gas chromatography coupled with time-of-flight mass spectrometry and partial least squares analysis. Analytical and Bioanalytical Chemistry, 407(1), 321–330. https://doi.org/10.1007/s00216-014-8233-6

Kiefl, J., Cordero, C., Nicolotti, L., Schieberle, P., Reichenbach, S. E., & Bicchi, C. (2012). Performance evaluation of non-targeted peak-based cross-sample analysis for comprehensive two-dimensional gas chromatography-mass spectrometry data and application to processed hazelnut profiling. Journal of Chromatography A, 1243, 81–90. https://doi.org/10.1016/j.chroma.2012.04.048

Kim, S., Ouyang, M., Jeong, J., Shen, C., & Zhang, X. (2014). A new method of peak detection for analysis of comprehensive two-dimensional gas chromatography mass spectrometry data. The Annals of Applied Statistics, 8(2), 1209–1231. https://doi.org/10.1214/14-AOAS731

Kim, S., Jang, H., Koo, I., Lee, J., & Zhang, X. (2017). Normal–Gamma–Bernoulli peak detection for analysis of comprehensive two-dimensional gas chromatography mass spectrometry data. Computational Statistics & Data Analysis, 105, 96–111. https://doi.org/10.1016/j.csda.2016.07.015

Kim, S. J., Serrano, G., Wise, K. D., Kurabayashi, K., & Zellers, E. T. (2011). Evaluation of a microfabricated thermal modulator for comprehensive two-dimensional microscale gas chromatography. Analytical Chemistry, 83(14), 5556–5562. https://doi.org/10.1021/ac200336e

Kind, T., Wohlgemuth, G., Lee, D. Y., Lu, Y., Palazoglu, M., Shahbaz, S., & Fiehn, O. (2009). FiehnLib: Mass spectral and retention index libraries for metabolomics based on quadrupole and time-of-flight gas chromatography/mass spectrometry. Analytical Chemistry, 81(24), 10038–10048. https://doi.org/10.1021/ac9019522

Klee, M. S., Cochran, J., Merrick, M., & Blumberg, L. M. (2015). Evaluation of conditions of comprehensive two-dimensional gas chromatography that yield a near-theoretical

maximum in peak capacity gain. Journal of Chromatography A, 1383, 151–159. https://doi.org/10.1016/j.chroma.2015.01.031

Knothe, G., & Steidley, K. R. (2005). Kinematic viscosity of biodiesel fuel components and related compounds. Influence of compound structure and comparison to petrodiesel fuel components. Fuel, 84(9), 1059–1065. https://doi.org/10.1016/j.fuel.2005.01.016

Koek, M. M., Jellema, R. H., van der Greef, J., Tas, A. C., & Hankemeier, T. (2011). Quantitative metabolomics based on gas chromatography mass spectrometry: Status and perspectives. Metabolomics, 7(3), 307–328. https://doi.org/10.1007/s11306-010-0254-3

Kondo, E., Marriott, P. J., Parker, R. M., Kouremenos, K. A., Morrison, P., & Adams, M. (2014). Metabolic profiling of yeast culture using gas chromatography coupled with orthogonal acceleration accurate mass time-of-flight mass spectrometry: Application to biomarker discovery. Analytica Chimica Acta, 807, 135–142. https://doi.org/10.1016/j.aca.2013.11.004

Korytár, P., Leonards, P. E. G., De Boer, J., & Brinkman, U. A. T. (2005). Group separation of organohalogenated compounds by means of comprehensive two-dimensional gas chromatography. Journal of Chromatography A, 1086, 29–44. https://doi.org/10.1016/j.chroma.2005.05.087

Kotseridis, Y. S., Spink, M., Brindle, I. D., Blake, A. J., Sears, M., Chen, X., et al. (2008). Quantitative analysis of 3-alkyl-2-methoxypyrazines in juice and wine using stable isotope labelled internal standard assay. Journal of Chromatography A, 1190(1–2), 294–301. https://doi.org/10.1016/j.chroma.2008.02.088

Krupčík, J., Gorovenko, R., Špánik, I., Sandra, P., & Armstrong, D. W. (2013). Flow-modulated comprehensive two-dimensional gas chromatography with simultaneous flame ionization and quadrupole mass spectrometric detection. Journal of Chromatography A, 1280, 104–111. https://doi.org/10.1016/j.chroma.2013.01.015

Krupčík, J., Májek, P., Gorovenko, R., Špánik, I., Sandra, P., & Armstrong, D. W. (2013). On the determination of a detector response enhancement factor for flow modulated comprehensive two-dimensional gas chromatography. Journal of Chromatography A, 1286, 235–240. https://doi.org/10.1016/j.chroma.2013.02.068

Kuprowicz, N. J., Zabarnick, S., West, Z. J., & Ervin, J. S. (2007). Use of measured species class concentrations with chemical kinetic modeling for the prediction of autoxidation and deposition of jet fuels. Energy & Fuels, 21(2), 530–544. https://doi.org/10.1021/ef060391o

Kwon, B. C., Eysenbach, B., Verma, J., Ng, K., De Filippi, C., Stewart, W. F., & Perer, A. (2018). Clustervision: Visual Supervision of Unsupervised Clustering. IEEE Transactions on Visualization and Computer Graphics, 24(1), 142–151. https://doi.org/10.1109/TVCG.2017.2745085

Lashgari, M., Singh, V., & Pawliszyn, J. (2019). A critical review on regulatory sample preparation methods: Validating solid-phase microextraction techniques. TrAC - Trends in Analytical Chemistry, 119, 115618. https://doi.org/10.1016/j.trac.2019.07.029

Lebanov, L., Tedone, L., Ghiasvand, A., & Paull, B. (2020). Random Forests machine learning applied to gas chromatography – Mass spectrometry derived average mass spectrum data

sets for classification and characterisation of essential oils. Talanta, 208(October 2019), 120471. https://doi.org/10.1016/j.talanta.2019.120471

Lee, A. L., Bartle, K. D., & Lewis, A. C. (2001). A model of peak amplitude enhancement in orthogonal two-dimensional gas chromatography. Analytical Chemistry, 73, 1330–1335. https://doi.org/10.1021/ac001120s

Lee, J., Flores-Cerrillo, J., Wang, J., & He, Q. P. (2020). A Variable Selection Method for Improving Variable Selection Consistency and Soft Sensor Performance. Proceedings of the American Control Conference, 2020-July, 725–730. https://doi.org/10.23919/ACC45564.2020.9147774

Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R. P., Tang, J., & Liu, H. (2017). Feature selection: A data perspective. ACM Computing Surveys, 50(6). https://doi.org/10.1145/3136625

Liang, J., Bai, L., Dang, C., & Cao, F. (2012). The K-type algorithms versus imbalanced data distributions. IEEE Transactions on Fuzzy Systems, 20(4), 728–745. https://doi.org/10.1109/TFUZZ.2011.2182354

Liu, S., & Davis, J. M. (2006). Dependence on saturation of average minimum resolution in two-dimensional statistical-overlap theory: Peak overlap in saturated two-dimensional separations. Journal of Chromatography A, 1126, 244–256. https://doi.org/10.1016/j.chroma.2006.05.064

Liu, Z., & Phillips, J. B. (1991). Comprehensive two-dimensional gas chromatography using an on-column thermal modulator interface. Journal of Chromatographic Science, 29, 227–231. https://doi.org/10.1093/chromsci/29.6.227

Liu, Z., & Phillips, J. B. (1994). Sensitivity and detection limit enhancement of gas chromatographic detection by thermal modulation. Journal of Microcolumn Separations, 6(3), 229–235. https://doi.org/10.1002/mcs.1220060306

Liu, Z., Patterson, D. G., & Lee, M. L. (1995). Geometric approach to factor analysis for the estimation of orthogonality and practical peak capacity in comprehensive two-dimensional separations. Analytical Chemistry, 67(21), 3840–3845. https://doi.org/10.1021/ac00117a004

Loohach, R., & Garg, K. (2012). Effect of Distance Functions on Simple K-means Clustering Algorithm. International Journal of Computer Applications, 49(6), 7–9. https://doi.org/10.5120/7629-0698

Lopatka, M., Sampat, A. A., Jonkers, S., Adutwum, L. A., Mol, H. G. J., van der Weg, G., et al. (2017). Local Ion Signatures (LIS) for the examination of comprehensive two-dimensional gas chromatography applied to fire debris analysis. Forensic Chemistry, 3, 1–13. https://doi.org/10.1016/j.forc.2016.10.003

Lopes, G. R., Petronilho, S., Ferreira, A. S., Pinto, M., Passos, C. P., Coelho, E., et al. (2021). Insights on Single-Dose Espresso Coffee Capsules' Volatile Profile: From Ground Powder Volatiles to Prediction of Espresso Brew Aroma Properties. Foods, 10(10), 2508. https://doi.org/10.3390/foods10102508

López-Galilea, I., Fournier, N., Cid, C., & Guichard, E. (2006). Changes in headspace volatile concentrations of coffee brews caused by the roasting process and the brewing procedure. Journal of Agricultural and Food Chemistry, 54(22), 8560–8566. https://doi.org/10.1021/jf061178t

Lovestead, T. M., Windom, B. C., Riggs, J. R., Nickell, C., & Bruno, T. J. (2010). Assessment of the compositional variability of RP-1 and RP-2 with the advanced distillation curve approach. Energy & Fuels, 24(10), 5611–5623. https://doi.org/10.1021/ef100994w

Lovestead, T. M., Burger, J. L., Schneider, N., & Bruno, T. J. (2016). Comprehensive Assessment of Composition and Thermochemical Variability by High Resolution GC/QToF-MS and the Advanced Distillation-Curve Method as a Basis of Comparison for Reference Fuel Development. Energy & Fuels, 30(12), 10029–10044. https://doi.org/10.1021/acs.energyfuels.6b01837

Luong, J., Guan, X., Xu, S., Gras, R., & Shellie, R. A. (2016). Thermal Independent Modulator for Comprehensive Two-Dimensional Gas Chromatography. Analytical Chemistry, 88(17), 8428–8432. https://doi.org/10.1021/acs.analchem.6b02525

MacQueen, J. B. (1967). Some Methods for Classification and Analysis of Multivariate Observations. In Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability (pp. 281–297).

De Maesschalck, R., Jouan-Rimbaud, D., & Massart, D. L. (2000). The Mahalanobis distance. Chemometrics and Intelligent Laboratory Systems, 50(1), 1–18. https://doi.org/10.1016/S0169-7439(99)00047-7

Magagna, F., Guglielmetti, A., Liberto, E., Reichenbach, S. E., Allegrucci, E., Gobino, G., et al. (2017). Comprehensive Chemical Fingerprinting of High-Quality Cocoa at Early Stages of Processing: Effectiveness of Combined Untargeted and Targeted Approaches for Classification and Discrimination. Journal of Agricultural and Food Chemistry, 65(30), 6329–6341. https://doi.org/10.1021/acs.jafc.7b02167

Malinowski, E. R. (1982). Obtaining the key set of typical vectors by factor analysis and subsequent isolation of component spectra. Analytica Chimica Acta, 134, 129–137. https://doi.org/10.1016/S0003-2670(01)84184-2

Malmquist, G., & Danielsson, R. (1994). Alignment of chromatographic profiles for principal component analysis: a prerequisite for fingerprinting methods. Journal of Chromatography A, 687(1), 71–88. https://doi.org/10.1016/0021-9673(94)00726-8

Mancha Agresti, P. D. C., Franca, A. S., Oliveira, L. S., & Augusti, R. (2008). Discrimination between defective and non-defective Brazilian coffee beans by their volatile profile. Food Chemistry, 106(2), 787–796. https://doi.org/10.1016/j.foodchem.2007.06.019

Marney, L. C., Siegler, W. C., Parsons, B. A., Hoggard, J. C., Wright, B. W., & Synovec, R. E. (2013). Tile-based Fisher-ratio software for improved feature selection analysis of comprehensive two-dimensional gas chromatography-time-of-flight mass spectrometry data. Talanta, 115, 887–895. https://doi.org/10.1016/j.talanta.2013.06.038

Mehmood, T., Liland, K. H., Snipen, L., & Sæbø, S. (2012). A review of variable selection methods in Partial Least Squares Regression. Chemometrics and Intelligent Laboratory Systems, 118, 62–69. https://doi.org/10.1016/j.chemolab.2012.07.010

Mikaliunaite, L., & Synovec, R. E. (2022). Computational method for untargeted determination of cycling yeast metabolites using comprehensive two-dimensional gas chromatography time-of-flight mass spectrometry. Talanta, 244, 123396. https://doi.org/10.1016/j.talanta.2022.123396

Van Mispelaar, V. G., Tas, A. C., Smilde, A. K., Schoenmakers, P. J., & Van Asten, A. C. (2003). Quantitative analysis of target components by comprehensive two-dimensional gas chromatography. Journal of Chromatography A, 1019, 15–29. https://doi.org/10.1016/j.chroma.2003.08.101

Mohan, A. R., & Eser, S. (2010). Analysis of carbonaceous solid deposits from thermal oxidative stressing of Jet-A fuel on iron- and nickel-based alloy surfaces. Industrial and Engineering Chemistry Research, 49(6), 2722–2730. https://doi.org/10.1021/ie901283r

Mohler, R. E., Dombek, K. M., Hoggard, J. C., Young, E. T., & Synovec, R. E. (2006). Comprehensive two-dimensional gas chromatography time-of-flight mass spectrometry analysis of metabolites in fermenting and respiring yeast cells. Analytical Chemistry, 78(8), 2700–2709. https://doi.org/10.1021/ac052106o

Mohler, R. E., Prazen, B. J., & Synovec, R. E. (2006). Total-transfer, valve-based comprehensive two-dimensional gas chromatography. Analytica Chimica Acta, 555(1), 68–74. https://doi.org/10.1016/j.aca.2005.08.072

Mohler, R. E., Dombek, K. M., Hoggard, J. C., Pierce, K. M., Young, E. T., & Synovec, R. E. (2007). Comprehensive analysis of yeast metabolite GC×GC-TOFMS data: Combining discovery-mode and deconvolution chemometric software. Analyst, 132(8), 756–767. https://doi.org/10.1039/b700061h

Mohler, R. E., O'Reilly, K. T., Zemo, D. A., Tiwary, A. K., Magaw, R. I., & Synowiec, K. A. (2013). Non-targeted analysis of petroleum metabolites in groundwater using GC×GC-TOFMS. Environmental Science and Technology, 47(18), 10471–10476. https://doi.org/10.1021/es401706m

Moreira de Oliveira, A., Alberto Teixeira, C., & Wang Hantao, L. (2022). Evaluation of the retention profile in flow-modulated comprehensive two-dimensional gas chromatography and independent component analysis of weathered heavy oils. Microchemical Journal, 172(PB), 106978. https://doi.org/10.1016/j.microc.2021.106978

Moros, G., Chatziioannou, A. C., Gika, H. G., Raikos, N., & Theodoridis, G. (2017). Investigation of the derivatization conditions for GC-MS metabolomics of biological samples. Bioanalysis, 9(1), 53–65. https://doi.org/10.4155/bio-2016-0224

Murphy, R. E., Schure, M. R., & Foley, J. P. (1998). Effect of Sampling Rate on Resolution in Comprehensive Two-Dimensional Liquid Chromatography. Analytical Chemistry, 70(8), 1585–1594. https://doi.org/10.1021/ac971184b

Murtada, K., Bowman, D., Edwards, M., & Pawliszyn, J. (2022). Thin-film microextraction combined with comprehensive two-dimensional gas chromatography time-of-flight mass spectrometry screening for presence of multiclass organic pollutants in drinking water samples. Talanta, 242, 123301. https://doi.org/10.1016/j.talanta.2022.123301

Muscalu, A. M., Edwards, M., Górecki, T., & Reiner, E. J. (2015). Evaluation of a single-stage consumable-free modulator for comprehensive two-dimensional gas chromatography:

Analysis of polychlorinated biphenyls, organochlorine pesticides and chlorobenzenes. Journal of Chromatography A, 1391(1), 93–101. https://doi.org/10.1016/j.chroma.2015.02.074

Mutarutwa, D., Navarini, L., Lonzarich, V., Crisafulli, P., Compagnone, D., & Pittia, P. (2020). Determination of 3-Alkyl-2-methoxypyrazines in Green Coffee: A Study to Unravel Their Role on Coffee Quality. Journal of Agricultural and Food Chemistry, 68(17), 4743–4751. https://doi.org/10.1021/acs.jafc.9b07476

Myers, O. D., Sumner, S. J., Li, S., Barnes, S., & Du, X. (2017). One Step Forward for Reducing False Positive and False Negative Compound Identifications from Mass Spectrometry Metabolomics Data: New Algorithms for Constructing Extracted Ion Chromatograms and Detecting Chromatographic Peaks. Analytical Chemistry, 89(17), 8696–8703. https://doi.org/10.1021/acs.analchem.7b00947

Navarro-Reig, M., Bedia, C., Tauler, R., & Jaumot, J. (2018). Chemometric strategies for peak detection and profiling from multidimensional chromatography. Proteomics, 18(18), 1–12. https://doi.org/10.1002/pmic.201700327

Ndayambaje, J. B., Nsabimana, A., Dushime, S., Ishimwe, F., Janvier, H., & Ongol, M. P. (2019). Microbial identification of potato taste defect from coffee beans. Food Science and Nutrition, 7(1), 287–292. https://doi.org/10.1002/fsn3.887

Novaes, F. J. M., Silva Junior, A. I. da, Kulsing, C., Nolvachai, Y., Bizzo, H. R., Aquino Neto, F. R. de, et al. (2019). New approaches to monitor semi-volatile organic compounds released during coffee roasting using flow-through/active sampling and comprehensive two-dimensional gas chromatography. Food Research International, 119(August 2018), 349–358. https://doi.org/10.1016/j.foodres.2019.02.009

Ochoa, G. S., Prebihalo, S. E., Reaser, B. C., Marney, L. C., & Synovec, R. E. (2020). Statistical inference of mass channel purity from Fisher ratio analysis using comprehensive two-dimensional gas chromatography with time of flight mass spectrometry data. Journal of Chromatography A, 1627, 461401. https://doi.org/10.1016/j.chroma.2020.461401

Ochoa, G. S., Billingsley, M. C., & Synovec, R. E. (2022). Using solid-phase extraction to facilitate a focused tile-based Fisher ratio analysis of comprehensive two-dimensional gas chromatography time-of-flight mass spectrometry data: comparative analysis of aerospace fuel composition. Analytical and Bioanalytical Chemistry, 415, 2411-2423. https://doi.org/10.1007/s00216-022-04348-1

Ochoa, G. S., Sudol, P. E., Trinklein, T. J., & Synovec, R. E. (2022). Class comparison enabled mass spectrum purification for comprehensive two-dimensional gas chromatography with time-of-flight mass spectrometry. Talanta, 236, 122844. https://doi.org/10.1016/j.talanta.2021.122844

de Oliveira Junqueira, A. C., de Melo Pereira, G. V., Coral Medina, J. D., Alvear, M. C. R., Rosero, R., de Carvalho Neto, D. P., et al. (2019). First description of bacterial and fungal communities in Colombian coffee beans fermentation analysed using Illumina-based amplicon sequencing. Scientific Reports, 9(1), 1–10. https://doi.org/10.1038/s41598-019-45002-8

Olivieri, A. C. (2021). A down-to-earth analyst view of rotational ambiguity in second-order calibration with multivariate curve resolution − a tutorial. Analytica Chimica Acta, 1156, 338206. https://doi.org/10.1016/j.aca.2021.338206

Omar, J., Olivares, M., Amigo, J. M., & Etxebarria, N. (2014). Resolution of co-eluting compounds of Cannabis Sativa in comprehensive two-dimensional gas chromatography/mass spectrometry detection with Multivariate Curve Resolution-Alternating Least Squares. Talanta, 121, 273–280. https://doi.org/10.1016/j.talanta.2013.12.044

Oros, F. J., & Davis, J. M. (1992). Comparison of statistical theories of spot overlap in two-dimensional separations and verification of means for estimating the number of zones. Journal of Chromatography, 591, 1–18.

Ortmayr, K., Charwat, V., Kasper, C., Hann, S., & Koellensperger, G. (2017). Uncertainty budgeting in fold change determination and implications for non-targeted metabolomics studies in model systems. Analyst, 142(1), 80–90. https://doi.org/10.1039/c6an01342b

Ott, L. S., Hadler, A. B., & Bruno, T. J. (2008). Variability of the rocket propellants RP-1, RP-2, and TS-5: Application of a composition- and enthalpy-explicit distillation curve method. Industrial and Engineering Chemistry Research, 47(23), 9225–9233. https://doi.org/10.1021/ie800988u

Outcalt, S. L., Laesecke, A., & Brumback, K. J. (2009). Thermophysical properties measurements of rocket propellants RP-1 and RP-2. Journal of Propulsion and Power, 25(5), 1032–1040. https://doi.org/10.2514/1.40543

Pani, O., & Górecki, T. (2006). Comprehensive two-dimensional gas chromatography (GC×GC) in environmental analysis and monitoring. Analytical and Bioanalytical Chemistry, 386(4), 1013–1023. https://doi.org/10.1007/s00216-006-0568-1

Panić, O., Górecki, T., McNeish, C., Goldstein, A. H., Williams, B. J., Worton, D. R., et al. (2011). Development of a new consumable-free thermal modulator for comprehensive two-dimensional gas chromatography. Journal of Chromatography A, 1218(20), 3070–3079. https://doi.org/10.1016/j.chroma.2011.03.024

Parastar, H., Radović, J. R., Jalali-Heravi, M., Diez, S., Bayona, J. M., & Tauler, R. (2011). Resolution and quantification of complex mixtures of polycyclic aromatic hydrocarbons in heavy fuel oil sample by means of GC × GC-TOFMS combined to multivariate curve resolution. Analytical Chemistry, 83(24), 9289–9297. https://doi.org/10.1021/ac201799r

Parsons, B. A., Marney, L. C., Siegler, W. C., Hoggard, J. C., Wright, B. W., & Synovec, R. E. (2015). Tile-Based Fisher Ratio Analysis of Comprehensive Two-Dimensional Gas Chromatography Time-of-Flight Mass Spectrometry (GC × GC–TOFMS) Data Using a Null Distribution Approach. Analytical Chemistry, 87(7), 3812–3819. https://doi.org/10.1021/ac504472s

Parsons, B. A., Pinkerton, D. K., Wright, B. W., & Synovec, R. E. (2016). Chemical characterization of the acid alteration of diesel fuel: Non-targeted analysis by two-dimensional gas chromatography coupled with time-of-flight mass spectrometry with tile-based Fisher ratio and combinatorial threshold determination. Journal of Chromatography A, 1440, 179–190. https://doi.org/10.1016/j.chroma.2016.02.067

Pasadakis, N., Gidarakos, E., Kanellopoulou, G., & Spanoudakis, N. (2008). Identifying sources of oil spills in a refinery by gas chromatography and chemometrics: A case study. Environmental Forensics, 9(1), 33–39. https://doi.org/10.1080/15275920701729548

Patchava, K. C., Benaissa, M., & Behairy, H. (2015). Improving the prediction performance of PLSR using RReliefF and FSD for the quantitative analysis of glucose in Near Infrared spectra. Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS, (1), 2379–2382. https://doi.org/10.1109/EMBC.2015.7318872

Peikari, M., Salama, S., Nofech-Mozes, S., & Martel, A. L. (2018). A Cluster-then-label Semi-supervised Learning Approach for Pathology Image Classification. Scientific Reports, 8(1), 1–13. https://doi.org/10.1038/s41598-018-24876-0

Pellegrino Vidal, R. B., Allegrini, F., & Olivieri, A. C. (2018). The effect of constraints on the analytical figures of merit achieved by extended multivariate curve resolution-alternating least-squares. Analytica Chimica Acta, 1003, 10–15. https://doi.org/10.1016/j.aca.2017.12.008

Pellegrino Vidal, R. B., Olivieri, A. C., & Tauler, R. (2018). Quantifying the Prediction Error in Analytical Multivariate Curve Resolution Studies of Multicomponent Systems. Analytical Chemistry, 90(11), 7040–7047. https://doi.org/10.1021/acs.analchem.8b01431

Peters, S., Vivó-Truyols, G., Marriott, P. J., & Schoenmakers, P. J. (2007). Development of an algorithm for peak detection in comprehensive two-dimensional chromatography. Journal of Chromatography A, 1156, 14–24. https://doi.org/10.1016/j.chroma.2006.10.066

Pierce, K. M., Hope, J. L., Johnson, K. J., Wright, B. W., & Synovec, R. E. (2005). Classification of gasoline data obtained by gas chromatography using a piecewise alignment algorithm combined with feature selection and principal component analysis. Journal of Chromatography A, 1096, 101–110. https://doi.org/10.1016/j.chroma.2005.04.078

Pierce, K. M., & Schale, S. P. (2011). Predicting percent composition of blends of biodiesel and conventional diesel using gas chromatography-mass spectrometry, comprehensive two-dimensional gas chromatography-mass spectrometry, and partial least squares analysis. Talanta, 83(4), 1254–1259. https://doi.org/10.1016/j.talanta.2010.07.084

Pinkerton, D. K., Parsons, B. A., Anderson, T. J., & Synovec, R. E. (2015). Trilinearity deviation ratio: A new metric for chemometric analysis of comprehensive two-dimensional gas chromatography time-of-flight mass spectrometry data. Analytica Chimica Acta, 871, 66–76. https://doi.org/10.1016/j.aca.2015.02.040

Pinkerton, D. K., Reaser, B. C., Berrier, K. L., & Synovec, R. E. (2017). Determining the probability of achieving a successful quantitative analysis for gas chromatography-mass spectrometry. Analytical Chemistry, 89(18), 9926–9933. https://doi.org/10.1021/acs.analchem.7b02230

Pizarro, C., Esteban-Díez, I., Sáenz-González, C., & González-Sáiz, J. M. (2008). Vinegar classification based on feature extraction and selection from headspace solid-phase microextraction/gas chromatography volatile analyses: A feasibility study. Analytica Chimica Acta, 608(1), 38–47. https://doi.org/10.1016/j.aca.2007.12.006

Poisson, L., Blank, I., Dunkel, A., & Hofmann, T. (2017). The Chemistry of Roasting—Decoding Flavor Formation. In The Craft and Science of Coffee (pp. 273–309). London, UK: Elsevier. https://doi.org/10.1016/B978-0-12-803520-7.00012-8

Pollo, B. J., Alexandrino, G. L., Augusto, F., & Hantao, L. W. (2018). The impact of comprehensive two-dimensional gas chromatography on oil & gas analysis: Recent advances and applications in petroleum industry. TrAC - Trends in Analytical Chemistry, 105, 202–217. https://doi.org/10.1016/j.trac.2018.05.007

Prebihalo, S. E., Berrier, K. L., Freye, C. E., Bahaghighat, H. D., Moore, N. R., Pinkerton, D. K., & Synovec, R. E. (2018). Multidimensional Gas Chromatography: Advances in Instrumentation, Chemometrics, and Applications. Analytical Chemistry, 90(1), 505–532. https://doi.org/10.1021/acs.analchem.7b04226

Prebihalo, S. E., Pinkerton, D. K., & Synovec, R. E. (2019). Impact of comprehensive two-dimensional gas chromatography time-of-flight mass spectrometry experimental design on data trilinearity and parallel factor analysis deconvolution. Journal of Chromatography A, 1605, 460368. https://doi.org/10.1016/j.chroma.2019.460368

Prebihalo, S. E., Ochoa, G. S., Berrier, K. L., Skogerboe, K. J., Cameron, K. L., Trump, J. R., et al. (2020). Control-Normalized Fisher Ratio Analysis of Comprehensive Two-Dimensional Gas Chromatography Time-of-Flight Mass Spectrometry Data for Enhanced Biomarker Discovery in a Metabolomic Study of Orthopedic Knee-Ligament Injury. Analytical Chemistry, 92(23), 15526–15533. https://doi.org/10.1021/acs.analchem.0c03456

Pua, A., Huang, Y., Vivian Goh, R. M., Ee, K.-H., Li, L., Cornuz, M., et al. (2023). Multidimensional Gas Chromatography of Organosulfur Compounds in Coffee and Structure–Odor Analysis of 2-Methyltetrahydrothiophen-3-one. Journal of Agricultural and Food Chemistry, 71(10), 4337–4345. https://doi.org/10.1021/acs.jafc.2c08842

Purcaro, G., Stefanuto, P. H., Franchina, F. A., Beccaria, M., Wieland-Alter, W. F., Wright, P. F., & Hill, J. E. (2018). SPME-GC×GC-TOFMS fingerprint of virally-infected cell culture: Sample preparation optimization and data processing evaluation. Analytica Chimica Acta, 1027, 158–167. https://doi.org/10.1016/j.aca.2018.03.037

Ralston-Hooper, K., Hopf, A., Oh, C., Zhang, X., Adamec, J., & Sepúlveda, M. S. (2008). Development of GCxGC/TOF-MS metabolomics for use in ecotoxicological studies with invertebrates. Aquatic Toxicology, 88(1), 48–52. https://doi.org/10.1016/j.aquatox.2008.03.002

Rasmussen, L. G., Savorani, F., Larsen, T. M., Dragsted, L. O., Astrup, A., & Engelsen, S. B. (2011). Standardization of factors that influence human urine metabolomics. Metabolomics, 7(1), 71–83. https://doi.org/10.1007/s11306-010-0234-7

Reaser, B. C., Wright, B. W., & Synovec, R. E. (2017). Using Receiver Operating Characteristic Curves to Optimize Discovery-Based Software with Comprehensive Two-Dimensional Gas Chromatography with Time-of-Flight Mass Spectrometry. Analytical Chemistry, 89(6), 3606–3612. https://doi.org/10.1021/acs.analchem.6b04991

Reichenbach, S. E., Ni, M., Kottapalli, V., & Visvanathan, A. (2004). Information technologies for comprehensive two-dimensional gas chromatography. Chemometrics and Intelligent Laboratory Systems, 71(2), 107–120. https://doi.org/10.1016/j.chemolab.2003.12.009

Reichenbach, S. E., Tian, X., Tao, Q., Stoll, D. R., & Carr, P. W. (2010). Comprehensive feature analysis for sample classification with comprehensive two-dimensional LC. Journal of Separation Science, 33(10), 1365–1374. https://doi.org/10.1002/jssc.200900859

Ren, S., Hinzman, A. A., Kang, E. L., Szczesniak, R. D., & Lu, L. J. (2015). Computational and statistical analysis of metabolomics data. Metabolomics, 11(6), 1492–1513. https://doi.org/10.1007/s11306-015-0823-6

Ribeiro, J. S., Augusto, F., Salva, T. J. G., Thomaziello, R. A., & Ferreira, M. M. C. (2009). Prediction of sensory properties of Brazilian Arabica roasted coffees by headspace solid phase microextraction-gas chromatography and partial least squares. Analytica Chimica Acta, 634(2), 172–179. https://doi.org/10.1016/j.aca.2008.12.028

Robnik-Šikonja, M., & Kononenko, I. (1997). An adaptation of Relief for attribute estimation in regression. Machine Learning: Proceedings of the Fourteenth International Conference (ICML'97), 5, 296–304.

Robnik-Šikonja, M., & Kononenko, I. (2003). Theoretical and Empirical Analysis of ReliefF and RReliefF. Machine Learning, 53, 23–69. https://doi.org/https://doi.org/10.1023/A:1025667309714

Rohatgi, V. K., & Ehsanes Saleh, A. K. M. (2015). An Introduction to Probability and Statistics (Third Edit). Hoboken, NJ: John Wiley & Sons.

Rousseau, R., Govaerts, B., Verleysen, M., & Boulanger, B. (2008). Comparison of some chemometric tools for metabonomics biomarker identification. Chemometrics and Intelligent Laboratory Systems, 91(1), 54–66. https://doi.org/10.1016/j.chemolab.2007.06.008

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics, 20(C), 53–65. https://doi.org/10.1016/0377-0427(87)90125-7

Ruckebusch, C., & Blanchet, L. (2013). Multivariate curve resolution: A review of advanced and tailored applications and challenges. Analytica Chimica Acta, 765, 28–36. https://doi.org/10.1016/j.aca.2012.12.028

Rutan, S. C., de Juan, A., & Tauler, R. (2009). Introduction to Multivariate Curve Resolution. In S. D. Brown, R. Tauler, & B. Walczak (Eds.), Comprehensive Chemometrics (Vol. 2, pp. 249–259). Elsevier.

Ryan, D., Shellie, R., Tranchida, P., Casilli, A., Mondello, L., & Marriott, P. (2004). Analysis of roasted coffee bean volatiles by using comprehensive two-dimensional gas chromatography-time-of-flight mass spectrometry. Journal of Chromatography A, 1054(1–2), 57–65. https://doi.org/10.1016/j.chroma.2004.08.057

Sampat, A. A. S., Daelen, B. Van, Lopatka, M., Mol, H., Derweg, G. Van, Truyols, G. V., et al. (2018). Detection and characterization of ignitable liquid residues in forensic fire debris

samples by comprehensive two-dimensional gas chromatography. Separations, 5(3), 1–27. https://doi.org/10.3390/separations5030043

Santos, F. J., & Galceran, M. T. (2003). Modern developments in gas chromatography-mass spectrometry-based environmental analysis. Journal of Chromatography A, 1000, 125–151. https://doi.org/10.1016/S0021-9673(03)00305-4

Savareear, B., Brokl, M., Wright, C., & Focant, J. F. (2017). Thermal desorption comprehensive two-dimensional gas chromatography coupled to time of flight mass spectrometry for vapour phase mainstream tobacco smoke analysis. Journal of Chromatography A, 1525, 126–137. https://doi.org/10.1016/j.chroma.2017.10.013

Scheidig, C., Czerny, M., & Schieberle, P. (2007). Changes in key odorants of raw coffee beans during storage under defined conditions. Journal of Agricultural and Food Chemistry, 55(14), 5768–5775. https://doi.org/10.1021/jf070488o

Schieberle, P., & Grosch, W. (1987). Quantitative Analysis of Aroma Compounds in Wheat and Rye Bread Crusts Using a Stable Isotope Dilution Assay. Journal of Agricultural and Food Chemistry, 35(2), 252–257. https://doi.org/10.1021/jf00074a021

Schmarr, H. G., & Bernhardt, J. (2010). Profiling analysis of volatile compounds from fruits using comprehensive two-dimensional gas chromatography and image processing techniques. Journal of Chromatography A, 1217(4), 565–574. https://doi.org/10.1016/j.chroma.2009.11.063

Schöneich, S., Cain, C. N., Freye, C. E., & Synovec, R. E. (2022). Optimization of Parameters for ROI Data Compression for Nontargeted Analyses Using LC–HRMS. Analytical Chemistry, 95(2), 1513–1521. https://doi.org/10.1021/acs.analchem.2c04538

Schöneich, S., Cain, C. N., Sudol, P. E., & Synovec, R. E. (2023). Enabling cuboid-based fisher ratio analysis using total-transfer comprehensive three-dimensional gas chromatography with time-of-flight mass spectrometry. Journal of Chromatography A, 1708, 464341. https://doi.org/10.1016/j.chroma.2023.464341

Schöneich, S., Gough, D. V, Trinklein, T. J., & Synovec, R. E. (2020). Dynamic pressure gradient modulation for comprehensive two-dimensional gas chromatography with time-of-flight mass spectrometry detection. Journal of Chromatography A, 1620, 460982. https://doi.org/https://doi.org/10.1016/j.chroma.2020.460982

Schöneich, S., Trinklein, T. J., Warren, C. G., & Synovec, R. E. (2020). A systematic investigation of comprehensive two-dimensional gas chromatography time-of-flight mass spectrometry with dynamic pressure gradient modulation for high peak capacity separations. Analytica Chimica Acta, 1134, 115–124. https://doi.org/10.1016/j.aca.2020.08.023

Schöneich, S., Ochoa, G. S., Monzón, C. M., & Synovec, R. E. (2022). Minimum variance optimized Fisher ratio analysis of comprehensive two-dimensional gas chromatography / mass spectrometry data: Study of the pacu fish metabolome. Journal of Chromatography A, 1667, 462868. https://doi.org/10.1016/j.chroma.2022.462868

Schostack, K. J., & Malinowski, E. R. (1993). Investigation of window factor analysis and matrix regression analysis in chromatography. Chemometrics and Intelligent Laboratory Systems, 20(2), 173–182. https://doi.org/10.1016/0169-7439(93)80013-8

Schure, M. R. (2011). The dimensionality of chromatographic separations. Journal of Chromatography A, 1218(2), 293–302. https://doi.org/10.1016/j.chroma.2010.11.016

Seeley, J. V. (2002). Theoretical study of incomplete sampling of the first dimension in comprehensive two-dimensional chromatography. Journal of Chromatography A, 962(1–2), 21–27. https://doi.org/10.1016/S0021-9673(02)00461-2

Seeley, J. V., Micyus, N. J., Bandurski, S. V., Seeley, S. K., & McCurry, J. D. (2007). Microfluidic deans switch for comprehensive two-dimensional gas chromatography. Analytical Chemistry, 79(5), 1840–1847. https://doi.org/10.1021/ac061881g

Seeley, J. V., Schimmel, N. E., & Seeley, S. K. (2018). The multi-mode modulator: A versatile fluidic device for two-dimensional gas chromatography. Journal of Chromatography A, 1536, 6–15. https://doi.org/10.1016/j.chroma.2017.06.030

Shellie, R. A., Welthagen, W., Zrostliková, J., Spranger, J., Ristow, M., Fiehn, O., & Zimmermann, R. (2005). Statistical methods for comparing comprehensive two-dimensional gas chromatography-time-of-flight mass spectrometry results: Metabolomic analysis of mouse tissue extracts. Journal of Chromatography A, 1086, 83–90. https://doi.org/10.1016/j.chroma.2005.05.088

Shen, X., Wang, B., Zi, C., Huang, L., Wang, Q., Zhou, C., et al. (2023). Interaction and Metabolic Function of Microbiota during the Washed Processing of Coffea arabica. Molecules, 28(16). https://doi.org/10.3390/molecules28166092

Shi, W., & Davis, J. M. (1993). Test of theory of overlap for two-dimensional separations by computer simulations of three-dimensional concentration profiles. Analytical Chemistry, 65(4), 482–492. https://doi.org/10.1021/ac00052a028

Shi, X., Li, H., Song, Z., Zhang, X., & Liu, G. (2017). Quantitative composition-property relationship of aviation hydrocarbon fuel based on comprehensive two-dimensional gas chromatography with mass spectrometry and flame ionization detector. Fuel, 200, 395–406. https://doi.org/10.1016/j.fuel.2017.03.073

Shingiro, J. B., Shee, P. K., Beaudry, R. M., Thiagarajan, D., Bourquin, L. D., & Walker, K. D. (2022). Assessing Alkyl Methoxypyrazines as Predictors of the Potato-Taste Defect in Coffee. ACS Food Science and Technology, 2(11), 1738–1745. https://doi.org/10.1021/acsfoodscitech.2c00233

Siegler, W. C., Fitz, B. D., Hoggard, J. C., & Synovec, R. E. (2011). Experimental study of the quantitative precision for valve-based comprehensive two-dimensional gas chromatography. Analytical Chemistry, 83(13), 5190–5196. https://doi.org/10.1021/ac200302b

Sinkov, N. A., & Harynuk, J. J. (2011). Cluster resolution: A metric for automated, objective and optimized feature selection in chemometric modeling. Talanta, 83(4), 1079–1087. https://doi.org/10.1016/j.talanta.2010.10.025

Skoczyńska, E., Korytár, P., & De Boer, J. (2008). Maximizing chromatographic information from environmental extracts by GCxGC-ToF-MS. Environmental Science and Technology, 42(17), 6611–6618. https://doi.org/10.1021/es703229t

Skov, T., Hoggard, J. C., Bro, R., & Synovec, R. E. (2009). Handling within run retention time shifts in two-dimensional chromatography data using shift correction and modeling. Journal of Chromatography A, 1216(18), 4020–4029. https://doi.org/10.1016/j.chroma.2009.02.049

Smith, S. W. (1999). Statistics, Probability and Noise. In The Scientist and Engineer's Guide to Digital Signal Processing Statistics, Probability and Noise (2nd ed., pp. 11–34). California Technical Publishing.

Sobkowiak, M., Griffith, J. M., Wang, B., & Beaver, B. (2009). Insight into the mechanisms of middle distillate fuel oxidative degradation. Part 1: On the role of phenol, indole, and carbazole derivatives in the thermal oxidative stability of fischer-tropsch/petroleum jet fuel blends. Energy & Fuels, 23(4), 2041–2046. https://doi.org/10.1021/ef8006992

Song, H., & Liu, J. (2018). GC-O-MS technique and its applications in food flavor analysis. Food Research International, 114, 187–198. https://doi.org/10.1016/j.foodres.2018.07.037

Song, S. M., Marriott, P., Kotsos, A., Drummer, O. H., & Wynne, P. (2004). Comprehensive two-dimensional gas chromatography with time-of-flight mass spectrometry (GC × GC-TOFMS) for drug screening and confirmation. Forensic Science International, 143, 87–101. https://doi.org/10.1016/j.forsciint.2004.02.042

Specialty Coffee Association of America. (2015). Cupping Specialty Coffee. Retrieved September 17, 2020, from http://www.scaa.org/PDF/resources/cupping-protocols.pdf

Stefanuto, P. H., Perrault, K. A., Stadler, S., Pesesse, R., Leblanc, H. N., Forbes, S. L., & Focant, J. F. (2015). GC×GC-TOFMS and supervised multivariate approaches to study human cadaveric decomposition olfactive signatures. Analytical and Bioanalytical Chemistry, 407(16), 4767–4778. https://doi.org/10.1007/s00216-015-8683-5

Stefanuto, P. H., Perrault, K. A., Dubois, L. M., L'Homme, B., Allen, C., Loughnane, C., et al. (2017). Advanced method optimization for volatile aroma profiling of beer using two-dimensional gas chromatography time-of-flight mass spectrometry. Journal of Chromatography A, 1507, 45–52. https://doi.org/10.1016/j.chroma.2017.05.064

Stein, S. E. (2012). Mass spectral reference libraries: An ever-expanding resource for chemical identification. Analytical Chemistry, 84(17), 7274–7282. https://doi.org/10.1021/ac301205z

Stein, S. E., & Scott, D. R. (1994). Optimization and testing of mass spectral library search algorithms for compound identification. Journal of the American Society for Mass Spectrometry, 5(9), 859–866. https://doi.org/10.1016/1044-0305(94)87009-8

Stilo, F., Bicchi, C., Robbat, A., Reichenbach, S. E., & Cordero, C. (2021). Untargeted approaches in food-omics: The potential of comprehensive two-dimensional gas chromatography/mass spectrometry. TrAC - Trends in Analytical Chemistry, 135, 116162. https://doi.org/10.1016/j.trac.2020.116162

van Stokkum, I. H. M., Mullen, K. M., & Mihaleva, V. V. (2009). Global analysis of multiple gas chromatography-mass spectrometry (GC/MS) data sets: A method for resolution of co-eluting components with comparison to MCR-ALS. Chemometrics and Intelligent Laboratory Systems, 95(2), 150–163. https://doi.org/10.1016/j.chemolab.2008.10.004

Strocchi, G., Bagnulo, E., Ruosi, M. R., Ravaioli, G., Trapani, F., Bicchi, C., et al. (2022). Potential Aroma Chemical Fingerprint of Oxidised Coffee Note by HS-SPME-GC-MS and Machine Learning. Foods, 11(24), 1–14. https://doi.org/10.3390/foods11244083

Sudol, P. E., Gough, D. V., Prebihalo, S. E., & Synovec, R. E. (2020). Impact of data bin size on the classification of diesel fuels using comprehensive two-dimensional gas chromatography with principal component analysis. Talanta, 206, 120239. https://doi.org/10.1016/j.talanta.2019.120239

Sudol, P. E., Pierce, K. M., Prebihalo, S. E., Skogerboe, K. J., Wright, B. W., & Synovec, R. E. (2020). Development of gas chromatographic pattern recognition and classification tools for compliance and forensic analyses of fuels: A review. Analytica Chimica Acta, 1132, 157–186. https://doi.org/10.1016/j.aca.2020.07.027

Sudol, P. E., Ochoa, G. S., & Synovec, R. E. (2021). Investigation of the limit of discovery using tile-based Fisher ratio analysis with comprehensive two-dimensional gas chromatography time-of-flight mass spectrometry. Journal of Chromatography A, 1644, 462092. https://doi.org/10.1016/j.chroma.2021.462092

Sudol, P. E., Galletta, M., Tranchida, P. Q., Zoccali, M., Mondello, L., & Synovec, R. E. (2022). Untargeted profiling and differentiation of geographical variants of wine samples using headspace solid-phase microextraction flow-modulated comprehensive two-dimensional gas chromatography with the support of tile-based Fisher ratio analysis. Journal of Chromatography A, 1662, 462735. https://doi.org/10.1016/j.chroma.2021.462735

Sudol, P. E., Ochoa, G. S., Cain, C. N., & Synovec, R. E. (2022). Tile-based variance rank initiated-unsupervised sample indexing for comprehensive two-dimensional gas chromatography-time-of-flight mass spectrometry. Analytica Chimica Acta, 1209, 339847. https://doi.org/10.1016/j.aca.2022.339847

Sunarharum, W. B., Williams, D. J., & Smyth, H. E. (2014). Complexity of coffee flavor: A compositional and sensory perspective. Food Research International, 62, 315–325. https://doi.org/10.1016/j.foodres.2014.02.030

Synovec, R. E., Freye, C. E., Billingsley, M. C., Keim, N., & Hill-Lam, B. (2018). Recent Advances in Relating Chemical Compositional Variation in RP-1, RP-2, and Similar Fuels to Thermal Integrity Data. In JANNAF 10th Liquid Propulsion Meeting. Long Beach, CA.

Synovec, R. E., Freye, C. E., Parsons, B. A., Billingsley, M. C., Keim, N., Hill-Lam, B., & Wilhelm, J. C. (2015). Recent Advances in the Development of Chemical Analysis Tools to Relate Compositional Variation to Thermal Integrity Data for RP-1, RP-2, and Related Fuels. In JANNAF 8th Liquid Propulsion Meeting. Nashville, TN.

Synovec, R. E., Freye, C. E., Billingsley, M. C., Keim, N., & Hill-Lam, B. (2016). Recent Advances in the Development of Chemical Analysis Tools to Relate Compositional Variation to Thermal Integrity Data for RP-1, RP-2, and Related Fuels. In JANNAF 9th Liquid Propulsion Meeting. Phoenix, AZ.

Tauler, R. (2001). Calculation of maximum and minimum band boundaries of feasible solutions for species profiles obtained by multivariate curve resolution. Journal of Chemometrics, 15(8), 627–646. https://doi.org/10.1002/cem.654

Tauler, R. (1995). Multivariate curve resolution applied to second order data. Chemometrics and Intelligent Laboratory Systems, 30(1), 133–146. https://doi.org/10.1016/0169-7439(95)00047-X

Taware, R., Taunk, K., Pereira, J. A. M., Shirolkar, A., Soneji, D., Câmara, J. S., et al. (2018). Volatilomic insight of head and neck cancer via the effects observed on saliva metabolites. Scientific Reports, 8(1), 17725. https://doi.org/10.1038/s41598-018-35854-x

Taylor, C. W., & Bowden, S. A. (2023). What about nitrogen? Using nitrogen as a carrier gas during the analysis of petroleum biomarkers by gas chromatography mass spectrometry. Journal of Chromatography A, 1697, 463989. https://doi.org/10.1016/j.chroma.2023.463989

The Good Scents Company. (2018). The Good Scents Company Information System. Retrieved from http://www.thegoodscentscompany.com/.

Thoumsin, C. (2019). Data Collection Methodology and Accurate Instance Rate Determination in Coffees With Potato Taste Defect (PTD). Retrieved from https://counterculturecoffee.com/wp-content/uploads/2020/04/CCC-PTD-Paper-Final.pdf

Tikunov, Y., Lommen, A., De Vos, C. H. R., Verhoeven, H. A., Bino, R. J., Hall, R. D., & Bovy, A. G. (2005). A novel approach for nontargeted data analysis for metabolomics. Large-scale profiling of tomato fruit volatiles. Plant Physiology, 139(3), 1125–1137. https://doi.org/10.1104/pp.105.068130

Toci, A. T., & Farah, A. (2008). Volatile compounds as potential defective coffee beans' markers. Food Chemistry, 108(3), 1133–1141. https://doi.org/10.1016/j.foodchem.2007.11.064

Toci, A. T., & Farah, A. (2014). Volatile fingerprint of Brazilian defective coffee seeds: Corroboration of potential marker compounds and identification of new low quality indicators. Food Chemistry, 153, 298–314. https://doi.org/10.1016/j.foodchem.2013.12.040

Tran, H. T. M., Vargas, C. A. C., Slade Lee, L., Furtado, A., Smyth, H., & Henry, R. (2017). Variation in bean morphology and biochemical composition measured in different genetic groups of arabica coffee (Coffea arabica L.). Tree Genetics and Genomes, 13(3). https://doi.org/10.1007/s11295-017-1138-8

Tran, T. C., & Marriott, P. J. (2007). Characterization of incense smoke by solid phase microextraction-Comprehensive two-dimensional gas chromatography (GC×GC). Atmospheric Environment, 41(27), 5756–5768. https://doi.org/10.1016/j.atmosenv.2007.02.030

Tranchida, P. Q., Purcaro, G., Visco, A., Conte, L., Dugo, P., Dawes, P., & Mondello, L. (2011). A flexible loop-type flow modulator for comprehensive two-dimensional gas chromatography. Journal of Chromatography A, 1218(21), 3140–3145. https://doi.org/10.1016/j.chroma.2010.11.082

Tranchida, P. Q., Sciarrone, D., Dugo, P., & Mondello, L. (2012). Heart-cutting multidimensional gas chromatography: A review of recent evolution, applications, and future prospects. Analytica Chimica Acta, 716, 66–75. https://doi.org/10.1016/J.ACA.2011.12.015

Tranchida, P. Q., Franchina, F. A., Zoccali, M., Bonaccorsi, I., Cacciola, F., & Mondello, L. (2013). A direct sensitivity comparison between flow-modulated comprehensive 2D and 1D GC in untargeted and targeted MS-based experiments. Journal of Separation Science, 36(17), 2746–2752. https://doi.org/10.1002/jssc.201300423

Tranchida, P. Q., Franchina, F. A., Dugo, P., & Mondello, L. (2014a). Flow-modulation low-pressure comprehensive two-dimensional gas chromatography. Journal of Chromatography A, 1372, 236–244.

Tranchida, P. Q., Franchina, F. A., Dugo, P., & Mondello, L. (2014b). Use of greatly-reduced gas flows in flow-modulated comprehensive two-dimensional gas chromatography-mass spectrometry. Journal of Chromatography A, 1359, 271–276.

Tranchida, P. Q., Salivo, S., Franchina, F. A., & Mondello, L. (2015). Flow-modulated comprehensive two-dimensional gas chromatography combined with a high-resolution time-of-flight mass spectrometer: A proof-of-principle study. Analytical Chemistry, 87(5), 2925–2930. https://doi.org/10.1021/ac5044175

Trinklein, T. J., Gough, D. V., Warren, C. G., Ochoa, G. S., & Synovec, R. E. (2020). Dynamic pressure gradient modulation for comprehensive two-dimensional gas chromatography. Journal of Chromatography A, 1609, 460488. https://doi.org/10.1016/j.chroma.2019.460488

Trinklein, T. J., Schöneich, S., Sudol, P. E., Warren, C. G., Gough, D. V., & Synovec, R. E. (2020). Total-transfer comprehensive three-dimensional gas chromatography with time-of-flight mass spectrometry. Journal of Chromatography A, 1634, 461654. https://doi.org/10.1016/j.chroma.2020.461654

Trinklein, T. J., Jiang, J., & Synovec, R. E. (2022). Profiling Olefins in Gasoline by Bromination Using GC×GC-TOFMS Followed by Discovery-Based Comparative Analysis. Analytical Chemistry, 94(26), 9407–9414. https://doi.org/10.1021/acs.analchem.2c01549

Trinklein, T. J., & Synovec, R. E. (2022). Simulating comprehensive two-dimensional gas chromatography mass spectrometry data with realistic run-to-run shifting to evaluate the robustness of tile-based Fisher ratio analysis. Journal of Chromatography A, 1677, 463321. https://doi.org/10.1016/j.chroma.2022.463321

Trinklein, T. J., Cain, C. N., Ochoa, G. S., Schöneich, S., Mikaliunaite, L., & Synovec, R. E. (2023). Recent Advances in GC×GC and Chemometrics to Address Emerging Challenges in Nontargeted Analysis. Analytical Chemistry, 95, 264–286. https://doi.org/10.1021/acs.analchem.2c04235

Urbanowicz, R. J., Meeker, M., La Cava, W., Olson, R. S., & Moore, J. H. (2018). Relief-based feature selection: Introduction and review. Journal of Biomedical Informatics, 85(January), 189–203. https://doi.org/10.1016/j.jbi.2018.07.014

Ventura, G. T., Simoneit, B. R. T., Nelson, R. K., & Reddy, C. M. (2012). The composition, origin and fate of complex mixtures in the maltene fractions of hydrothermal petroleum assessed by comprehensive two-dimensional gas chromatography. Organic Geochemistry, 45, 48–65. https://doi.org/10.1016/j.orggeochem.2012.01.002

Vial, J., Noçairi, H., Sassiat, P., Mallipatu, S., Cognon, G., Thiébaut, D., et al. (2009). Combination of dynamic time warping and multivariate analysis for the comparison of

comprehensive two-dimensional gas chromatograms. Application to plant extracts. Journal of Chromatography A, 1216(14), 2866–2872. https://doi.org/10.1016/j.chroma.2008.09.027

Vivó-Truyols, G. (2012). Bayesian approach for peak detection in two-dimensional chromatography. Analytical Chemistry, 84(6), 2622–2630. https://doi.org/10.1021/ac202124t

Voigtman, E., & Winefordner, J. D. (1987). The Multiplex Disadvantage and Excess Low-Frequency Noise. Applied Spectroscopy, 41(7), 1182–1184. https://doi.org/10.1366/0003702874447509

Vozka, P., Mo, H., Šimáček, P., & Kilaz, G. (2018). Middle distillates hydrogen content via GC×GC-FID. Talanta, 186, 140–146. https://doi.org/10.1016/j.talanta.2018.04.059

Vozka, P., Modereger, B. A., Park, A. C., Zhang, W. T. J., Trice, R. W., Kenttämaa, H. I., & Kilaz, G. (2019). Jet fuel density via GC × GC-FID. Fuel, 235, 1052–1060. https://doi.org/10.1016/j.fuel.2018.08.110

Walczak, B., & Massart, D. L. (2001). Dealing with missing data: Part I. Chemometrics and Intelligent Laboratory Systems, 58(1), 15–27. https://doi.org/10.1016/S0169-7439(01)00131-9

Wang, B., Fang, A., Heim, J., Bogdanov, B., Pugh, S., Libardoni, M., & Zhang, X. (2010). DISCO: Distance and spectrum correlation optimization alignment for two-dimensional gas chromatography time-of-flight mass spectrometry-based metabolomics. Analytical Chemistry, 82(12), 5069–5081. https://doi.org/10.1021/ac100064b

Wang, Y., O'Reilly, J., Chen, Y., & Pawliszyn, J. (2005). Equilibrium in-fibre standardisation technique for solid-phase microextraction. Journal of Chromatography A, 1072(1), 13–17. https://doi.org/10.1016/j.chroma.2004.12.084

Wang, Z., Fingas, M., & Sigouin, L. (2002). Using multiple criteria for fingerprinting unknown oil samples having very similar chemical composition. Environmental Forensics, 3, 251–262. https://doi.org/10.1006/enfo.2002.0098

Watson, N. E., van Wingerden, M. M., Pierce, K. M., Wright, B. W., & Synovec, R. E. (2006). Classification of high-speed gas chromatography-mass spectrometry data by principal component analysis coupled with piecewise alignment and feature selection. Journal of Chromatography A, 1129(1), 111–118. https://doi.org/10.1016/j.chroma.2006.06.087

Watson, N. E., Parsons, B. A., & Synovec, R. E. (2016). Performance evaluation of tile-based Fisher Ratio analysis using a benchmark yeast metabolome dataset. Journal of Chromatography A, 1459, 101–111. https://doi.org/10.1016/j.chroma.2016.06.067

Webb-Robertson, B. J., Kim, Y. M., Zink, E. M., Hallaian, K. A., Zhang, Q., Madupu, R., et al. (2014). A statistical analysis of the effects of urease pre-treatment on the measurement of the urinary metabolome by gas chromatography-mass spectrometry. Metabolomics, 10(5), 897–908. https://doi.org/10.1007/s11306-014-0642-1

Webster, R. L., Rawson, P. M., Kulsing, C., Evans, D. J., & Marriott, P. J. (2017). Investigation of the Thermal Oxidation of Conventional and Alternate Aviation Fuels with Comprehensive Two-Dimensional Gas Chromatography Accurate Mass Quadrupole

Time-of-Flight Mass Spectrometry. Energy & Fuels, 31(5), 4886–4894. https://doi.org/10.1021/acs.energyfuels.7b00178

Weingart, G., Kluger, B., Forneck, A., Krska, R., & Schuhmacher, R. (2012). Establishment and application of a metabolomics workflow for identification and profiling of volatiles from leaves of Vitis vinifera by HS-SPME-GC-MS. Phytochemical Analysis, 23(4), 345–358. https://doi.org/10.1002/pca.1364

Widegren, J. A., & Bruno, T. J. (2013). Thermal stability of RP-2 as a function of composition: The effect of linear, branched, and cyclic alkanes. Energy & Fuels, 27(9), 5138–5143. https://doi.org/10.1021/ef401677g

Wilson, R. B., Siegler, W. C., Hoggard, J. C., Fitz, B. D., Nadeau, J. S., & Synovec, R. E. (2011). Achieving high peak capacity production for gas chromatography and comprehensive two-dimensional gas chromatography by minimizing off-column peak broadening. Journal of Chromatography A, 1218(21), 3130–3139. https://doi.org/10.1016/j.chroma.2010.12.108

Windig, W., & Guilment, J. (1991). Interactive Self-Modeling Mixture Analysis. Analytical Chemistry, 63, 1425–1432.

Windig, W., Bogomolov, A., & Kucheryavskiy, S. (2020). Two-Way Data Analysis: Detection of Purest Variables. In Comprehensive Chemometrics: Chemical and Biochemical Data Analysis (Second Edi, Vol. 2, pp. 107–136). Elsevier. https://doi.org/10.1016/B978-0-12-409547-2.14747-X

Wold, S., Esbensen, K., & Geladi, P. (1987). Principal component analysis. Chemometrics and Intelligent Laboratory Systems, 2(1–3), 37–52. https://doi.org/10.1016/0169-7439(87)80084-9

Wold, S., Sjöström, M., & Eriksson, L. (2001). PLS-regression: A basic tool of chemometrics. Chemometrics and Intelligent Laboratory Systems, 58(2), 109–130. https://doi.org/10.1016/S0169-7439(01)00155-1

Won, S. H., Veloo, P. S., Dooley, S., Santner, J., Haas, F. M., Ju, Y., & Dryer, F. L. (2016). Predicting the global combustion behaviors of petroleum-derived and alternative jet fuels by simple fuel property measurements. Fuel, 168, 34–46. https://doi.org/10.1016/j.fuel.2015.11.026

Worley, B., Halouska, S., & Powers, R. (2013). Utilities for quantifying separation in PCA/PLS-DA scores plots. Analytical Biochemistry, 433(2), 102–104. https://doi.org/10.1016/j.ab.2012.10.011

Yan, D. D., Wong, Y. F., Tedone, L., Shellie, R. A., Marriott, P. J., Whittock, S. P., & Koutoulis, A. (2018). Chemotyping of new hop (Humulus lupulus L.) genotypes using comprehensive two-dimensional gas chromatography with quadrupole accurate mass time-of-flight mass spectrometry. Journal of Chromatography A, 1536, 110–121. https://doi.org/10.1016/j.chroma.2017.08.020

Yang, N., Liu, C., Liu, X., Degn, T. K., Munchow, M., & Fisk, I. (2016). Determination of volatile marker compounds of common coffee roast defects. Food Chemistry, 211, 206–214. https://doi.org/10.1016/j.foodchem.2016.04.124

Yang, S., Hoggard, J. C., Lidstrom, M. E., & Synovec, R. E. (2013). Comprehensive discovery of 13C labeled metabolites in the bacterium Methylobacterium extorquens AM1 using gas chromatography-mass spectrometry. Journal of Chromatography A, 1317, 175–185. https://doi.org/10.1016/j.chroma.2013.08.059

Yu, J., & Eser, S. (1997). Thermal Decomposition of C10-C14 Normal Alkanes in Near-Critical and Supercritical Regions: Product Distributions and Reaction Mechanisms. Industrial and Engineering Chemistry Research, 36(3), 574–584. https://doi.org/10.1021/ie960392b

Yuan, C., Emelianov, D. A., Varfolomeev, M. A., & Abaas, M. (2019). Comparison of oxidation behavior of linear and branched alkanes. Fuel Processing Technology, 188(February), 203–211. https://doi.org/10.1016/j.fuproc.2019.02.025

Zabarnick, S., West, Z. J., Shafer, L. M., Mueller, S. S., Striebich, R. C., & Wrzesinski, P. J. (2019). Studies of the Role of Heteroatomic Species in Jet Fuel Thermal Stability: Model Fuel Mixtures and Real Fuels. Energy & Fuels, 33(9), 8557–8565. https://doi.org/10.1021/acs.energyfuels.9b02345

Zanella, D., Henin, A., Mascrez, S., Stefanuto, P. H., Franchina, F. A., Focant, J. F., & Purcaro, G. (2022). Comprehensive two-dimensional gas chromatographic platforms comparison for exhaled breath metabolites analysis. Journal of Separation Science, 45(18), 3542–3555. https://doi.org/10.1002/jssc.202200164

Zhang, D., Huang, X., Regnier, F. E., & Zhang, M. (2008). Two-dimensional correlation optimized warping algorithm for aligning GCxGC-MS data. Analytical Chemistry, 80(8), 2664–2671. https://doi.org/10.1021/ac7024317

Zhang, W., & Zhao, P. X. (2014). Quality evaluation of extracted ion chromatograms and chromatographic peaks in liquid chromatography/mass spectrometry-based metabolomics data. BMC Bioinformatics, 15(11), 1–13. https://doi.org/10.1186/1471-2105-15-S11-S5

Zhang, X., Zhang, Z., & Tauler, R. (2019). Evaluation of the extension of rotation ambiguity associated to multivariate curve resolution solutions by the application of the MCR-BANDS method. Talanta, 202(May), 554–564. https://doi.org/10.1016/j.talanta.2019.05.002

Zhang, Z., & Pawliszyn, J. (1993). Headspace Solid-Phase Microextraction. Analytical Chemistry, 65(14), 1843–1852. https://doi.org/10.1021/ac00062a008

Zhao, F., Wang, P., Lucardi, R. D., Su, Z., & Li, S. (2020). Natural sources and bioactivities of 2,4-di-tert-butylphenol and its analogs. Toxins, 12(1), 1–26. https://doi.org/10.3390/toxins12010035

Ziółkowska, A., Wąsowicz, E., & Jeleń, H. H. (2016). Differentiation of wines according to grape variety and geographical origin based on volatiles profiling using SPME-MS and SPME-GC/MS methods. Food Chemistry, 213, 714–720. https://doi.org/10.1016/j.foodchem.2016.06.120

Zou, Y., Gaida, M., Franchina, F. A., Stefanuto, P. H., & Focant, J. F. (2022). Distinguishing between Decaffeinated and Regular Coffee by HS-SPME-GC×GC-TOFMS, Chemometrics, and Machine Learning. Molecules, 27(6). https://doi.org/10.3390/molecules27061806

Zushi, Y., Gros, J., Tao, Q., Reichenbach, S. E., Hashimoto, S., & Arey, J. S. (2017). Pixel-by-pixel correction of retention time shifts in chromatograms from comprehensive two-dimensional gas chromatography coupled to high resolution time-of-flight mass spectrometry. Journal of Chromatography A, 1508, 121–129. https://doi.org/10.1016/j.chroma.2017.05.065

Appendix A

This appendix is reproduced from the Electronic Supplementary Material of C. N. Cain, N. J. Haughn, H. J. Purcell, L. C. Marney, R. E. Synovec, C. T. Thoumsin, S. C. Jackels, K. J. Skogerboe, Analytical Determination of the Severity of Potato Taste Defect in Roasted East African Arabica Coffee, J. Agric. Food Chem. 7 (2021) 2253-2261.

Table A.1. IPMP concentration and odor rankings for all 49 coffee samples analyzed.

Sample Number	IPMP Concentration (ng/g)	Odor Attribution
1	0.6	Clean
2	15.9	Mild
3	30.7	Medium
4	33.7	Strong
5	52.7	Strong
6	22.5	Strong
7	1.8	Clean
8	2.9	Clean
9	1.8	Clean
10	11.3	Mild
11	72.4	Mild
12	11.8	Mild
13	32.7	Medium
14	44.5	Medium
15	49.0	Medium
16	1.8	Clean
17	1.8	Clean
18	1.6	Mild
19	12.1	Medium
20	22.5	Medium
21	65.3	Strong
22	18.6	Strong
23	85.9	Strong
24	1.9	Clean
25	2.2	Clean
26	2.2	Clean
27	1.5	Clean
28	3.1	Clean
29	2.4	Clean
30	16.9	Mild
31	4.4	Mild

32	9.6	Mild
33	4.3	Mild
34	5.7	Medium
35	8.4	Medium
36	25.1	Medium
37	4.9	Medium
38	27.0	Medium
39	79.8	Medium
40	21.6	Strong
41	52.3	Strong
42	4.1	Strong
43	25.8	Strong
44	63.8	Strong
45	26.3	Strong
46	13.7	Strong
47	461.1	Strong
48	1.7	Clean
49	529.9	Strong

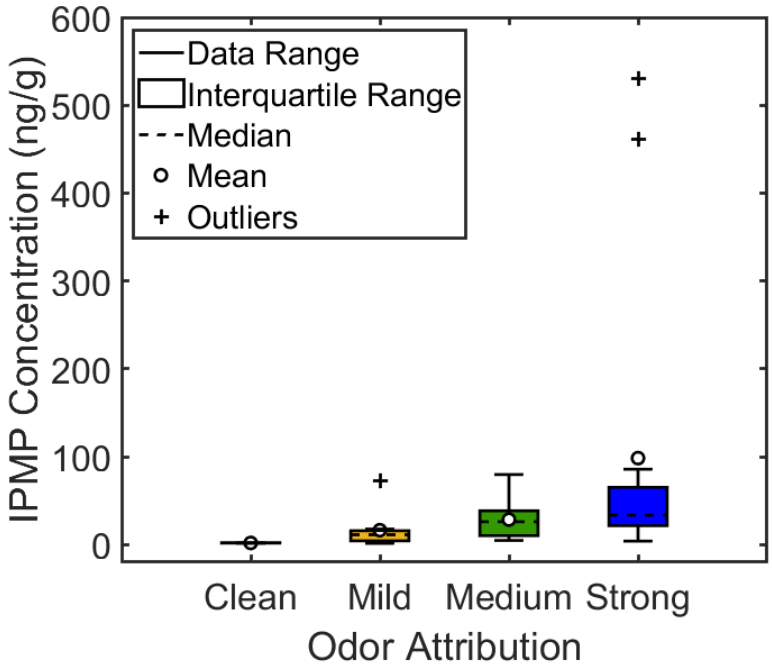


Figure A.1. Box-and-whisker plots relating IPMP concentration to the intensity of odor attributed to PTD before statistically removing outliers.

Table A.2. Summary statistics for IPMP concentration in each odor attribution category prior to removal of outliers.

Odor Attribution	Number of Samples	Range (ng/g)	Median (ng/g)	Average (ng/g)	Standard Error (ng/g)
Clean	13	0.6 - 3.1	1.8	2.0	0.2
Mild	9	1.6 - 72.4	11.3	16.5	7.2
Medium	12	4.9 - 79.8	26.1	28.5	6.2
Strong	15	4.1 - 529.9	33.7	98.5	42.2

Table A.3. Hit list for the clean versus strong PTD comparison using the traditional F-ratio calculation. The average F-ratio was found by taking the mean of the F-ratios from the top 3 *m/z*. The largest F-ratio and its corresponding *m/z* is also provided. For each hit and *m/z*, the average peak area in the strong PTD samples was divided by the average peak area in clean samples to produce a concentration ratio ([Strong]/[Clean] = [S]/[C]). Each hit shaded in blue showed a statistical difference in the peak areas of the clean and strong PTD classes with a *t*-test at the 95 % confidence interval.

Hit #	Avg. F-ratio	*m/z*	Largest F-ratio	t_R (min)	Tentative Compound Identification	MV	[S]/[C]	*p*-value
1	22.9	137	25.9	42.5	IPMP	936	21.4	0.001
2	13.5	110	18.7	49.3	5-Methyl-2-furancarboxaldehyde	901	0.91	0.194
3	12.9	91	13.4	57.5	Methyl salicylate	891	1.75	0.010
4	12	122	18	60.4	2-Methoxyphenol	800	2.16	0.027
5	11.1	80	11.3	27.4	1-Methyl-1H-pyrrole	832	0.63	0.007
6	11.1	65	12.4	73.6	Ethyl 4-ethoxybenzoate	803	1.95	0.013
7	10.4	123	13.5	40.4	2-Ethyl-5-methylpyrazine	867	0.80	0.042
8	10.4	108	14.2	38.6	2,3-Dimethylpyrazine	843	0.72	0.045
9	9.93	96	10.6	67.2	1H-Pyrrole-2-carboxaldehyde	837	1.42	0.004
10	9.78	96	12.2	44.5	Tetramethyl pyrazine	876	0.99	0.193
11	9.66	78	10.4	34.1	Styrene	800	0.95	0.330
12	9.54	108	10.2	38	Ethyl pyrazine	852	0.71	0.047
13	9.47	81	10.5	28	3-Methylphenol	832	0.74	0.020
14	9.1	126	9.73	45.5	Furfuryl formate	807	0.36	0.015
15	9.01	123	12.1	41.3	UnkA		1.12	0.229
16	8.26	63	11.6	60.3	3-Phenylfuran	820	1.71	0.046
17	8.15	83	9.5	59.1	1-Furfurylpyrrole	894	1.71	0.033
18	8.06	123	9.8	40.7	2-Ethyl-6-methylpyrazine	860	1.15	0.146
19	7.89	121	8.83	64.6	2-Acetylpyrrole	805	2.30	0.026
20	7.72	69	10.4	58.8	UnkB		1.02	0.052
21	7.66	95	7.8	57.3	Unk1		1.74	0.044

22	7.58	79	9.42	51	UnkC		0.98	0.071
23	7.33	94	8.3	34.5	Methyl pyrazine	890	0.89	0.214
24	7.33	95	10.3	60.5	UnkD		1.18	0.141
25	7.13	128	9.06	45.4	3,5-Diethyl-2-methylpyrazine	800	0.77	0.014
26	7.07	66	10.9	56	o-Methoxybenzenethiol	800	1.21	0.603
27	7.01	121	12.3	62.1	UnkE		1.05	0.089
28	6.93	82	8.74	64.9	Difurfuryl ether	820	2.15	0.045
29	6.56	59	8.66	53.8	1,6-Dihydrocarveol	801	1.65	0.026
30	6.55	123	7.78	66.1	UnkF		0.89	0.083
31	6.44	68	9.07	61.1	UnkG		1.19	0.181
32	6.1	51	7.44	50.4	2,2'-Difurylmethane	909	1.03	0.248
33	5.7	120	6.24	51.6	UnkH		0.88	0.064
34	5.69	51	6.48	47.2	Furfuryl acetate	828	0.92	0.311
35	5.4	109	5.93	64.2	UnkI		0.95	0.071
36	5.14	93	6.99	54.2	UnkJ		0.95	0.108
37	5	109	6.23	43.1	3-Ethyl-2,5-dimethylpyrazine	916	0.91	0.246
38	4.84	110	5.92	46.4	Benzofuran	800	1.06	0.798
39	4.67	92	4.94	68.7	UnkK		1.12	0.094
40	4.43	164	4.44	67	4-Ethyl-2-methoxyphenol	843	1.96	0.032
41	4.38	51	5.38	50	2,2'-Bifuran	851	0.85	0.165

Table A.4. Hit list for the clean versus strong PTD comparison using the clean-normalized F-ratio calculation. The average F-ratio was found by taking the mean of the F-ratios from the top 3 *m/z*. The largest F-ratio and its corresponding *m/z* is also provided. For each hit and *m/z*, the average peak area in the strong PTD samples was divided by the average peak area in clean samples to produce a concentration ratio ([Strong]/[Clean] = [S]/[C]). Each hit shaded in blue showed a statistical difference in the peak areas of the clean and strong PTD classes with a *t*-test at the 95 % confidence interval.

Hit #	Avg. F-ratio	*m/z*	Largest F-ratio	t_R (min)	Tentative Compound Identification	MV	[S]/[C]	*p*-value
1	8270	137	12400	42.5	IPMP	936	21.4	0.001
2	47.7	73	81.5	74.5	UnkL		1.13	0.220
3	31.3	126	64.5	74.3	4-Vinyl guaiacol	896	2.46	0.008
4	29.7	107	68.5	35.6	2,7-Dimethyloxepine	825	0.92	0.559
5	25.4	110	36.3	49.4	5-Methyl-2-furancarboxaldehyde	901	0.91	0.194
6	24.2	91	28.1	57.5	Methyl salicylate	891	1.75	0.010
7	21.9	121	46.9	62.1	UnkE		1.05	0.089
8	21.5	83	25.1	59.0	1-Furfurylpyrrole	894	1.71	0.033

9	18.7	122	26.7	60.4	2-Methoxyphenol	800	2.16	0.027
10	18.6	65	19.8	73.6	Ethyl 4-ethoxybenzoate	803	1.95	0.013
11	18.6	121	26.1	64.6	2-Acetylpyrrole	805	2.30	0.026
12	17.7	93	22.8	54.2	UnkJ		0.95	0.108
13	16.9	96	18.4	67.2	1H-Pyrrole-2-carboxaldehyde	837	1.42	0.004
14	16.1	69	20.3	58.8	UnkB		1.02	0.052
15	13.9	123	23.9	66.1	UnkF		0.89	0.083
16	13.8	157	17.6	56.2	UnkM		0.95	0.083
17	13.1	63	16.9	60.3	3-Phenylfuran	820	1.71	0.046
18	12.2	92	15.3	68.7	UnkK		1.12	0.094
19	11.3	78	12.7	34.1	Styrene	800	0.95	0.330
20	10.7	152	11.3	60.6	UnkN		1.09	0.065
21	9.61	96	14.1	44.5	Tetramethyl pyrazine	876	0.99	0.193
22	9.19	68	12.8	61.1	UnkG		1.19	0.181
23	8.81	123	10.2	40.4	2-Ethyl-5-methylpyrazine	867	0.8	0.042
24	8.78	80	8.83	27.4	1-Methyl-1H-pyrrole	832	0.63	0.007
25	8.55	59	12.1	53.8	1,6-Dihydrocarveol	801	1.65	0.026
26	8.15	140	8.68	53.0	UnkO		1.15	0.180
27	7.87	81	8.28	28.0	3-Methylphenol	832	0.74	0.020
28	7.67	164	8.60	67.0	4-Ethyl-2-methoxyphenol	843	1.96	0.032
29	7.47	82	9.37	64.9	Difurfuryl ether	820	2.15	0.045
30	7.32	79	11.0	51.0	UnkC		0.98	0.071
31	7.15	108	9.33	38.6	2,3-Dimethylpyrazine	843	0.72	0.045
32	7.14	95	7.70	57.3	Unk1		1.74	0.044
33	7.13	109	7.59	64.2	UnkI		0.95	0.071
34	7.11	150	7.84	66.2	UnkP		0.98	0.092
35	6.66	106	8.25	47.3	Benzaldehyde	870	1.16	0.612
36	6.55	108	7.81	38.0	Ethyl pyrazine	852	0.71	0.047
37	6.18	53	8.61	71.4	1-Methyl-1H-pyrrole-2-carboxaldehyde	801	0.89	0.080
38	6.01	66	6.28	65.9	Phenol	807	1.15	0.264
39	6.00	95	6.59	60.5	UnkD		1.18	0.141
40	5.95	51	6.61	50.4	2,2'-Difurylmethane	909	1.03	0.248
41	5.79	121	7.68	41.4	2-Ethyl-3-methylpyrazine	875	0.74	0.048
42	5.57	81	11.8	32.9	Furfuryl methyl ether	927	0.85	0.603
43	5.32	126	6.04	36.9	Methyl furoate	801	0.91	0.483
44	5.25	120	6.38	52.3	2-Furanmethanol	918	1.05	0.195
45	5.14	126	5.38	45.6	Furfuryl formate	807	0.36	0.015
46	5.12	120	6.06	51.6	UnkH		0.88	0.064
47	5.10	123	5.37	55.5	UnkQ		1.13	0.278

48	4.69	128	6.22	45.3	3,5-Diethyl-2-methylpyrazine	800	0.77	0.014
49	4.64	123	8.11	40.8	2-Ethyl-6-methylpyrazine	860	1.15	0.146

Table A.5. Hit list for the clean versus strong PTD comparison using the strong-normalized F-ratio calculation. The average F-ratio was found by taking the mean of the F-ratios from the top 3 *m/z*. The largest F-ratio and its corresponding *m/z* is also provided. For each hit and *m/z*, the average peak area in the strong PTD samples was divided by the average peak area in clean samples to produce a concentration ratio ([Strong]/[Clean] = [S]/[C]). Each hit shaded in blue showed a statistical difference in the peak areas of the clean and strong PTD classes with a *t*-test at the 95 % confidence interval.

Hit #	Avg. F-ratio	*m/z*	Largest F-ratio	t_R (min)	Tentative Compound Identification	MV	[S]/[C]	*p*-value
1	44	126	51	45.6	Furfuryl formate	807	0.36	0.015
2	37.8	128	54.7	45.4	3,5-Diethyl-2-methylpyrazine	800	0.77	0.014
3	25.0	108	30.1	38.7	2,3-Dimethylpyrazine	843	0.72	0.045
4	23.9	123	26.8	40.4	2-Ethyl-5-methylpyrazine	867	0.8	0.042
5	22.5	93	34.3	28.3	β-Myrcene	885	0.95	0.132
6	20.7	81	23.1	28.1	3-Methylphenol	832	0.74	0.020
7	20.0	78	20.5	34.1	Styrene	800	0.95	0.330
8	19.6	79	22.2	29.9	Pyridine	828	1.06	0.706
9	18.6	108	23.7	38.0	Ethyl pyrazine	852	0.71	0.047
10	17.1	123	18.7	41.3	UnkA		1.12	0.229
11	16.3	126	33.4	74.3	4-Vinyl guaiacol	896	2.46	0.008
12	15.1	66	34.1	56.0	o-Methoxybenzenethiol	800	1.21	0.603
13	15.1	80	16.0	27.4	1-Methyl-1H-pyrrole	832	0.63	0.007
14	15.1	109	19.7	43.1	3-Ethyl-2,5-dimethylpyrazine	916	0.91	0.246
15	14.7	51	16.1	47.3	Furfuryl acetate	828	0.92	0.311
16	13.0	94	15.3	34.5	Methyl pyrazine	890	0.89	0.214
17	11.9	96	13.1	44.5	Tetramethyl pyrazine	876	0.99	0.193
18	11.5	137	13.0	42.5	IPMP	936	21.4	0.001
19	10.9	108	15.4	37.7	2,6-Dimethylpyrazine	906	0.85	0.151
20	10.8	121	19.4	64.6	2-Acetylpyrrole	805	2.30	0.026
21	10.5	108	16.2	37.4	2,5-Dimethylpyrazine	909	0.97	0.157
22	9.83	79	15.9	51.0	UnkC		0.98	0.071
23	9.80	91	19.0	57.5	Methyl salicylate	891	1.75	0.010
24	9.67	123	11.4	40.7	2-Ethyl-6-methylpyrazine	860	1.15	0.146
25	9.59	108	10.17	43.9	2-Ethyl-3,5-dimethylpyrazine	861	0.88	0.167

26	9.48	122	15.9	60.4	2-Methoxyphenol	800	2.16	0.027
27	9.19	110	12.62	49.4	5-Methyl-2-furancarboxaldehyde	901	0.91	0.194
28	9.10	81	9.16	35.4	UnkR		0.95	0.142
29	8.83	95	10.82	57.3	Unk1		1.74	0.044
30	7.88	65	9.00	73.6	Ethyl 4-ethoxybenzoate	803	1.95	0.013
31	7.86	51	8.5	50.4	2,2'-Difurylmethane	909	1.03	0.248
32	7.65	129	9.87	73.0	UnkS		0.93	0.220
33	7.20	96	7.48	67.2	1H-Pyrrole-2-carboxaldehyde	837	1.42	0.004
34	7.12	120	9.08	51.6	UnkH		0.88	0.064
35	7.01	63	8.82	60.3	3-Phenylfuran	820	1.71	0.046
36	6.66	152	9.73	60.6	UnkN		1.09	0.065
37	6.64	135	11.89	42.7	2,6-Diethylpyrazine	855	0.8	0.044
38	6.59	82	8.20	64.9	Difurfuryl ether	820	2.15	0.045
39	6.48	135	6.57	39.0	UnkT		1.09	0.107
40	6.41	110	6.67	46.4	Benzofuran	800	1.06	0.798
41	6.40	69	18.66	58.8	UnkB		1.02	0.052
42	6.27	83	6.43	59.1	1-Furfurylpyrrole	894	1.71	0.033
43	5.74	59	6.7	53.8	1,6-Dihydrocarveol	801	1.65	0.026
44	4.99	51	6.21	50.0	2,2'-Bifuran	851	0.85	0.165
45	4.83	95	5.08	60.5	UnkD		1.18	0.141
46	4.55	109	10.07	48.0	UnkU		1.07	0.384
47	4.44	123	4.50	53.5	UnkV		1.13	0.128
48	4.35	93	4.39	54.2	UnkJ		0.95	0.108

Table A.6. The average peak area (± standard deviation) measured at the top F-ratio *m/z* for each analyte of interest (Table 2.2) in each PTD odor attribution class. The standard error of each measurement is also provided. A one-way ANOVA determined that the peak areas for each analyte were statistically different among the four classes at the 95 % confidence interval.

Compound	Peak Area				*p*-value
	Clean	Mild PTD	Medium PTD	Strong PTD	
Furfuryl formate	3.47 ± 0.74	2.22 ± 0.70	1.30 ± 0.31	1.26 ± 0.38	0.020
1-Methyl-1H-pyrrole	1.07 ± 0.10	0.65 ± 0.06	0.77 ± 0.08	0.68 ± 0.07	0.003
Ethyl pyrazine	10.4 ± 0.62	8.72 ± 1.09	8.76 ± 0.88	7.33 ± 0.51	0.042
2,3-Dimethylpyrazine	4.06 ± 0.38	4.33 ± 0.52	2.93 ± 0.36	2.92 ± 0.25	0.020
3-Methylphenol	0.76 ± 0.05	0.49 ± 0.09	0.54 ± 0.05	0.56 ± 0.03	0.003
2-Ethyl-3-methylpyrazine	1.43 ± 0.11	1.42 ± 0.10	1.16 ± 0.10	1.05 ± 0.11	0.037
3,5-Diethyl-2-methylpyrazine	2.67 ± 0.18	1.95 ± 0.23	2.35 ± 0.22	2.06 ± 0.09	0.043
2,6-Diethylpyrazine	6.53 ± 0.39	6.32 ± 0.32	5.22 ± 0.49	5.21 ± 0.42	0.048
2-Ethyl-5-methylpyrazine	1.41 ± 0.08	1.12 ± 0.11	1.06 ± 0.11	1.13 ± 0.08	0.034
1H-Pyrrole-2-carboxaldehyde	0.58 ± 0.04	0.69 ± 0.05	0.74 ± 0.06	0.83 ± 0.06	0.013
1,6-Dihydrocarveol	1.61 ± 0.10	1.47 ± 0.06	1.76 ± 0.03	2.47 ± 0.46	0.047
3-Phenylfuran	0.41 ± 0.04	0.45 ± 0.04	0.57 ± 0.03	0.64 ± 0.11	0.045
1-Furfurylpyrrole	0.52 ± 0.03	0.60 ± 0.04	0.70 ± 0.05	0.84 ± 0.13	0.039
Unk3	9.37 ± 0.63	10.7 ± 0.76	11.7 ± 0.66	15.4 ± 2.61	0.047
Methyl salicylate	0.88 ± 0.10	1.12 ± 0.08	1.28 ± 0.08	1.51 ± 0.14	0.001
Ethyl 4-ethoxybenzoate	0.94 ± 0.12	1.01 ± 0.08	1.56 ± 0.10	1.72 ± 0.38	0.049
4-Ethyl-2-methoxyphenol	1.28 ± 0.13	1.57 ± 0.15	2.20 ± 0.29	2.37 ± 0.44	0.040
Difurfuryl ether	1.59 ± 0.13	1.83 ± 0.11	2.83 ± 0.29	3.17 ± 0.70	0.032
2-Methoxyphenol	0.85 ± 0.09	1.10 ± 0.08	1.54 ± 0.18	1.72 ± 0.34	0.025
2-Acetylpyrrole	0.42 ± 0.06	0.56 ± 0.04	0.85 ± 0.09	0.98 ± 0.18	0.004
4-Vinyl guaiacol	0.71 ± 0.13	0.87 ± 0.08	1.53 ± 0.30	1.65 ± 0.28	0.013
IPMP	0.15 ± 0.01	1.18 ± 0.43	1.95 ± 0.58	2.98 ± 0.61	0.001

Appendix B

This appendix is reproduced from the Electronic Supplementary Material of C. N. Cain, M. Gaida, P.-H. Stefanuto, J.-F. Focant, R. E. Synovec, S. C. Jackels, K. J. Skogerboe, Investigating Sensory-Classified Roasted Arabica Coffee with GC×GC-TOFMS and Chemometrics to Understand Potato Taste Defect, Microchem. J. 196 (2024), 109578.

Table B.1. List of the IPMP concentrations and PTD odor attribution for the 56 coffee samples analyzed. Our previous publication describes how the concentration of IPMP and PTD odor attribution were determined [1].

Sample Number	IPMP Concentration (ng/g)	Odor Attribution
1	0.6	Clean
2	15.9	Mild
3	30.7	Medium
4	33.7	Strong
5	52.7	Strong
6	22.5	Strong
7	1.8	Clean
8	2.9	Clean
9	1.8	Clean
10	11.3	Mild
11	72.4	Mild
12	11.8	Mild
13	32.7	Medium
14	44.5	Medium
15	49.0	Medium
16	1.8	Clean
17	1.8	Clean
18	1.6	Mild
19	12.1	Medium
20	22.5	Medium
21	65.3	Strong
22	18.6	Strong
23	85.9	Strong
24	1.9	Clean
25	2.2	Clean
26	2.2	Clean
27	1.5	Clean
28	3.1	Clean
29	2.4	Clean
30	16.9	Mild
31	4.4	Mild

32	9.6	Mild
33	4.3	Mild
34	5.7	Medium
35	8.4	Medium
36	25.1	Medium
37	4.9	Medium
38	27.0	Medium
39	79.8	Medium
40	21.6	Strong
41	52.3	Strong
42	4.1	Strong
43	25.8	Strong
44	63.8	Strong
45	26.3	Strong
46	13.7	Strong
47	461.1	Strong
48	1.7	Clean
49	529.9	Strong
50	0.0	Clean
51	88.7	Strong
52	48.7	Strong
53	4.8	Mild
54	3.9	Mild
55	199	Strong
56	25.5	Medium

[1] C.N. Cain, N.J. Haughn, H.J. Purcell, L.C. Marney, R.E. Synovec, C.T. Thoumsin, S.C. Jackels, K.J. Skogerboe, Analytical Determination of the Severity of Potato Taste Defect in Roasted East African Arabica Coffee, J. Agric. Food Chem. (2021). https://doi.org/10.1021/acs.jafc.1c00605.

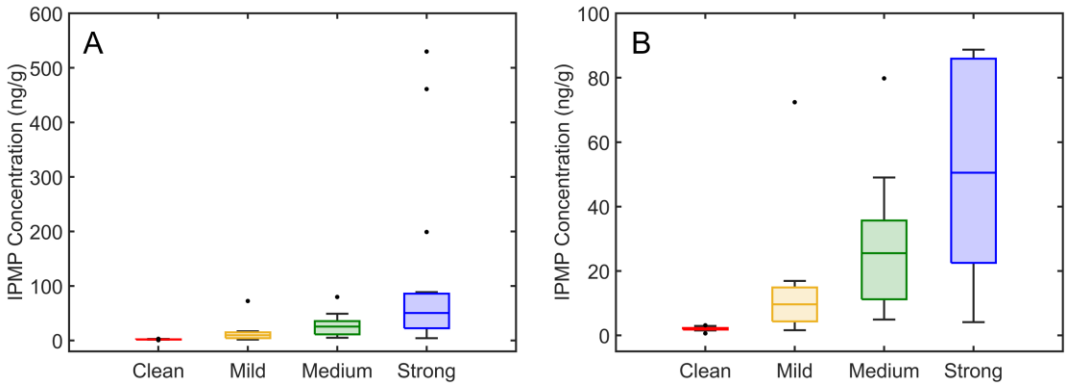


Figure B.1. (A) Box-and-whiskers plot of the IPMP concentration measured for all 56 coffee samples. (B) A zoom-in highlighting the IPMP concentration range from 0 ng/g to 100 ng/g. Samples with an outlier IPMP concentration are represented by a black dot.

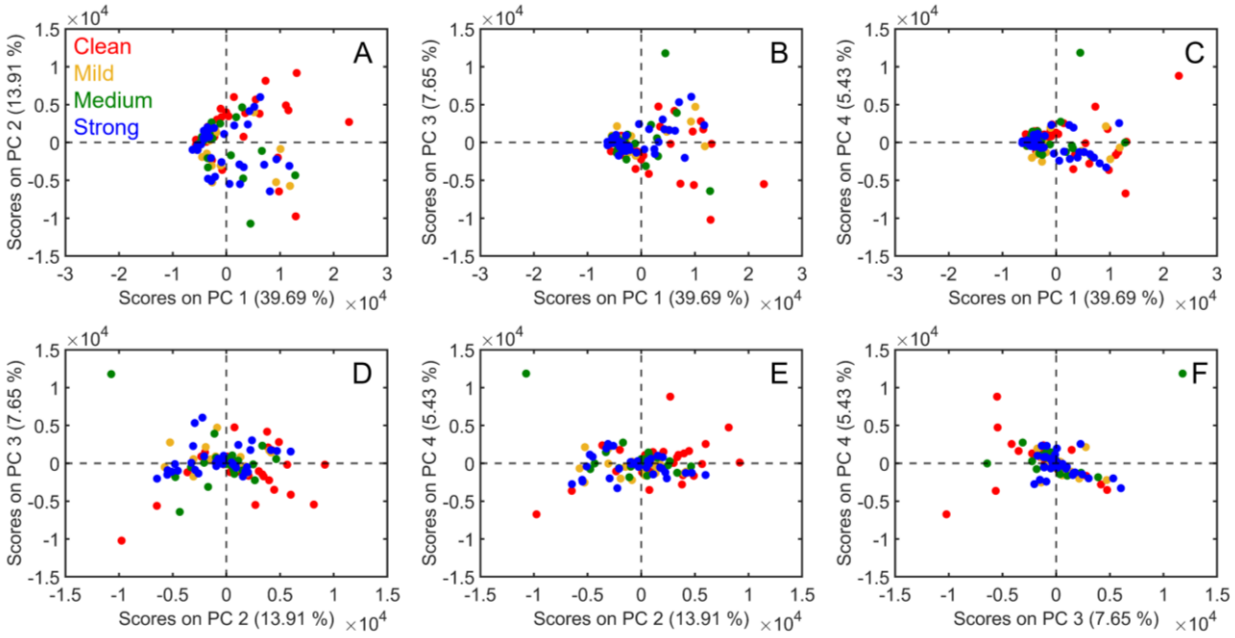


Figure B.2. Scores plot from PCA of the unfolded, normalized TIC chromatograms of the clean (red), mild PTD (yellow), medium PTD (green), and strong PTD (blue) coffee samples. Plots of (A) PC 2 versus PC 1, (B) PC 3 versus PC 1, (C) PC 4 versus PC 1, (D) PC 3 versus PC 2, (E) PC 4 versus PC 2, and (F) PC 4 versus PC 3 are provided.

Figure B.2 shows the resulting PCA scores plot using the unfolded, normalized TIC chromatograms for the clean (red), mild PTD (yellow), medium PTD (green), and strong PTD (blue) samples. Notably, the scores plots do not show any distinct clustering of the samples based on the presence of PTD. The lack of sample clustering on the scores plots are further illustrated by the fact that this PCA model only captures 66.67 % of the total variance in the first four PCs. As an unsupervised chemometric method, PCA is sensitive to the presence of instrumental artifacts like detector noise and retention time misalignment. Based on inspection of the chromatograms, retention time shifting was present in the data set, which hinders the ability of PCA to cluster these samples based on meaningful chemical differences.

Table B.2. List of all 359 class-distinguishing hits (p-value < 0.01) that were discovered using tile-based F-ratio analysis. The hit list is ranked in descending order of F-ratios. Tentative compound identifications were made if the mass spectrum match a library spectrum with a MV ≥ 800. Otherwise, peaks that could not be identified are listed as an unknown (Unk) and numbered. Concentration ratios for each analyte were calculated as [Strong]/[Clean] ([S]/[C] here) using a pure *m/z* [2]. The metrics for determining *m/z* purity (p-value and *LOF*) are also reported. Note, a *LOF* was not calculated when the analyte was present in only one class. The value for each hit (except IPMP) in the linear regression vector of the PLS model shown in Figure 3.7 is also provided. Sensory descriptions are listed for known analytes [3].

Hit #	^1t_R (min)	^2t_R (s)	F-ratio	Compound	MV	*m/z*	[S]/[C]	*p*	*LOF* (%)	LRV	Sensory
1	26.57	0.77	65.18	7-Benzofuranamine, 2-methyl-	811	147	0.55	7.5E-11	15	-4.08E-05	
2	18.67	1.47	54.3	Pyrazine, 2-methoxy-3-(1-methylethyl)-	849	152	20.4	4.9E-08	8.6	N/A	Earthy, Vegetable, Potato
3	43.23	0.97	52.95	Unk1		87	2.81	8.0E-04	7.7	4.80E-06	
4	38.3	1.16	49.39	1,3-Pentadiene, 1,1-diphenyl-, (Z)-	823	205	10.9	9.7E-08	15.7	2.86E-06	
5	25.03	0.86	45.61	3(2H)-Benzofuranone, 7-methyl-	801	148	0.47	4.7E-08	10.3	-6.20E-05	
6	39.77	1.43	43.54	Benzene, 1,1'-(1,1,2,2-tetramethyl-1,2-ethanediyl)bis-	817	119	20.7	1.8E-07	5.6	4.92E-06	
7	37.23	1.6	43.35	Benzene, 1,1'-(1,4-dimethyl-1-butene-1,4-diyl)bis-	809	221	S only	3.1E-07		1.39E-06	
8	37.83	1.29	43.03	1,5,6,7-Tetramethyl-3-phenylbicyclo[3.2.0]hepta-2,6-diene	803	194	15.1	2.6E-05	19.7	3.07E-06	
9	17.07	0.33	42.39	Pyridine, 3-ethyl-	813	136	0.31	5E-08	18.2	-3.78E-06	Tobacco, Caramel-like, Roasted, Hazelnut
10	22.8	0.94	41.67	Benzofuran, 2-methyl-	826	103	0.42	5.3E-07	12.7	-3.74E-05	Burnt, Phenolic
11	31.8	1.44	41.22	Unk2		99	5.17	2.5E-07	19.3	7.76E-06	
12	17.67	0.27	39.28	2,4,6-Octatriene, 2,6-dimethyl-	855	79	0.46	7.8E-09	15.6	-1.54E-05	Sweet, Floral, Nutty
13	23.7	0.87	39.03	1H-Indole, 2,3-dihydro-	801	117	0.49	1.9E-07	15.6	-5.00E-06	
14	18.57	0.97	39	Unk3		127	0.37	1.1E-07	4.9	-6.68E-06	
15	22.93	0.92	38.42	Unk4		77	0.53	5.6E-06	18.3	-4.61E-06	
16	22.63	0.93	38.24	2-Propenal, 3-phenyl-	869	133	0.17	5.9E-08	7.6	-9.84E-06	Sweet, Spicy, Honey, Cinnamon
17	14.4	1.02	37.84	Cyclohexene, 1-methyl-4-(1-methylethylidene)-	902	103	0.81	0.00031	16.8	-1.90E-06	Fresh, Sweet, Pine, Citrus

18	30.43	0.06	37.13	Unk5		110	S only	2.5E-06		1.05E-06	
19	37.53	1.33	36.17	Unk6		119	S only	1.2E-06		2.93E-06	
20	40.03	1.12	36.12	Unk7		222	S only	1.3E-06		8.89E-06	
21	21.07	1.2	36.06	3-Methyl-2,3-dihydro-benzofuran	803	105	0.39	2.5E-07	14.2	-2.88E-06	
22	22.2	1.2	35.45	Unk8		115	0.61	3.7E-07	15.3	-1.91E-06	
23	27.77	0.87	35.23	Unk9		152	0.56	0.00012	19.6	-4.90E-06	
24	35.77	1.37	35.07	1-Propene, 3-(2-cyclopentenyl)-2-methyl-1,1-diphenyl-	810	222	29.8	1.9E-06	10.8	1.34E-05	
25	17.97	1.34	34.44	2-Methyl-3-isopropylpyrazine	820	108	0.43	7E-06	10.7	-1.16E-05	Coffee
26	28.07	0.73	34.3	Unk10		117	0.5	1.7E-06	10.8	-1.26E-05	
27	22.67	1.16	34.1	Unk11		78	0.08	2.6E-07	12.6	-1.72E-06	
28	24.7	0.86	33.06	Furan, 2-(2-furanylmethyl)-5-methyl-	910	74	0.46	2.9E-07	9.1	-7.30E-06	
29	25.17	0.65	32.85	Unk12		150	0.36	1.6E-06	10.9	-6.99E-06	
30	20.17	1.13	32.78	Unk13		134	0.48	1.8E-06	13.2	-1.10E-06	
31	13.5	0.79	32.76	1,3,7-Octatriene, 3,7-dimethyl-	924	93	0.32	2.6E-06	14	-2.89E-06	Fruity, Floral
32	27.6	0.85	32.75	Acetic acid, 2-phenylethyl ester	802	159	0.53	1.1E-06	10	-1.05E-06	Floral, Sweet, Honey, Fruity, Cocoa
33	17.43	1.51	32.75	Pyrazine, 2-ethyl-5-methyl-	877	93	0.14	1.1E-07	11.6	-2.15E-06	Coffee, Nutty, Roasted
34	37.23	0.66	32.74	Phenol, 2,4-bis(1,1-dimethylethyl)-	908	191	4.5	2.5E-06	4	1.16E-03	
35	25.97	0.75	32.51	Unk14		126	0.39	7.8E-07	7.3	-5.61E-06	
36	21.4	1.52	32.43	cis-4-Decenal	809	98	0.41	3.5E-06	5.1	-2.76E-06	Citrus, Aldehydic, Cardamom
37	35.4	1.53	32.2	1H-Indene, 2,3-dihydro-1,1,3-trimethyl-3-phenyl-	896	236	19.6	6.8E-06	4.8	3.07E-06	
38	21.83	1.03	32.2	5,6,7,8-Tetrahydroquinoxaline	805	119	0.5	6.7E-07	19.5	-1.16E-06	Nutty, Roasted, Cereal
39	24.97	1.09	32.2	Benzofuran, 4,7-dimethyl-	862	144	0.57	2.9E-08	5.6	-4.10E-06	
40	38.9	1.31	31.75	2,4-Diphenyl-4-methyl-2(*E*)-pentene	883	236	19.8	4.3E-06	4.9	8.52E-05	
41	23.03	0.82	31.54	Furan, 2,2'-methylenebis-	901	100	0.47	6.8E-06	14.8	-8.37E-06	Rich, Roasted
42	17.83	1.24	31.21	3-Octen-2-one, (*E*)-	874	68	0.49	9.2E-07	18.4	-1.13E-05	
43	37.23	1.23	31.14	Benzene, (1,3-dimethyl-3-butenyl)-	807	105	S only	3.9E-06		1.14E-05	
44	27.07	0.92	31.13	Unk15		149	0.53	9.4E-07	8.1	-6.91E-06	

45	39.33	1.44	31.11	Benzene, [2-methyl-1-(1-methylethyl)propyl]-	801	91	S only	7.8E-07		2.11E-06	
46	39.43	1.14	30.83	Unk16		222	S only	1.1E-05		8.70E-07	
47	22.73	0.05	30.63	Unk17		121	0.54	2.1E-06	4.9	-1.04E-06	
48	24.9	0.81	30.56	Unk18		120	0.42	1.7E-07	14.4	-3.74E-05	
49	25.37	1.01	30.4	Unk19		146	0.62	6.7E-08	5.6	-2.89E-06	
50	18.53	1.05	30.06	Benzenethiol, 3-methyl-	802	125	0.27	3.9E-06	14.7	-8.70E-06	Burnt
51	28.57	0.75	29.82	4-Hydroxybenzo[b]thiophene	809	150	0.54	7.3E-08	6.4	-5.46E-05	
52	25.13	1.05	29.52	Unk20		154	0.53	3.2E-06	9	-8.14E-06	
53	20.7	1.11	29.5	Benzofuran, 2,3-dihydro-2-methyl-	800	65	0.4	2.7E-07	15	-9.27E-06	
54	26.03	0.64	29.43	1H-Indol-5-ol	848	88	0.4	1.1E-08	20	-4.24E-06	
55	23.37	0.83	29.43	m-Menth-1(7)-ene, (*R*)-(-)-	805	52	0.53	1.1E-05	18.6	-1.39E-06	
56	27.57	0.98	29.4	3-Acetyl-2,5-dimethylthiophene	819	85	0.55	7.1E-07	19.5	-6.42E-06	Burnt, Roasted, Nutty
57	22.27	1.05	29.39	2,2-Dimethylnon-5-en-3-one	810	132	0.1	1.8E-06	8.4	-2.50E-06	
58	25.87	0.76	29.32	2-Acetyl-3-methylthiophene	859	126	0.22	3.7E-09	5.7	-5.18E-06	Phenolic, Floral, Almond
59	18.77	1.19	29.2	Unk21		127	0.45	0.00012	14.7	-9.61E-07	
60	24.43	1.18	29.06	Unk22		105	0.47	1.6E-07	2.1	-6.99E-06	
61	26.67	1.8	28.84	Unk23		134	4.09	0.0001	9.2	5.89E-07	
62	19.17	1.92	28.45	Cyclodecanol	820	82	S only	5.3E-05		1.29E-06	
63	19.63	1.24	28.41	Pyrazine, tetramethyl-	873	140	0.46	3.1E-06	19.2	-8.93E-07	
64	22.47	1.24	28.31	Pyrazine, 2,3-dimethyl-5-(1-propenyl)-, (*E*)-	806	52	0.51	1.3E-06	19	-1.16E-06	
65	40.73	0.51	28.16	Unk24		60	0.59	6.8E-08	19	-3.47E-06	
66	19.93	0.88	28.08	Pyrazine, 2-ethenyl-6-methyl-	856	50	0.42	8.7E-06	11	-1.11E-06	Nutty, Hazelnut
67	21.1	0.89	27.75	Orcinol	801	154	0.52	1.8E-05	18.8	-2.91E-06	
68	16.2	1.63	27.64	Pyridine, 2,6-diethyl-	893	134	0.47	1.8E-06	15.9	-1.42E-06	
69	21	1.08	27.51	Unk25		98	0.36	8E-07	13.7	-1.42E-06	
70	21.27	1.05	27.51	Pyrazine, 2-methyl-6-(1-propenyl)-, (*Z*)-	834	52	0.44	3.9E-06	15.5	-9.28E-06	
71	17.73	1.52	27.42	Phenol, 2,4,5-trimethyl-	818	136	0.62	4.8E-05	18.8	-2.86E-06	Phenolic
72	19.1	1.54	27.23	Pyrazine, 2,5-dimethyl-3-propyl-	851	66	0.15	4.2E-06	12.8	-1.77E-06	Nutty, Hazelnut
73	24.43	0.93	27.13	1,3-Benzenediol, 4-propyl-	819	52	0.5	1.4E-06	15.8	-1.23E-05	

74	24.2	1.01	27.04	Unk26		106	0.53	3.2E-05	13.5	-1.95E-06	
75	27.27	0.97	26.99	Unk27		162	0.57	2.8E-06	3.3	-3.09E-06	
76	28.07	0.92	26.98	Unk28		162	0.58	2.4E-05	9	-2.04E-06	
77	18.83	1.32	26.97	Pyrazine, 3-ethyl-2,5-dimethyl-	886	128	0.41	5E-06	14.1	-7.89E-06	Cocoa, Roasted, Nutty
78	23.97	0.86	26.93	Unk29		148	0.43	4.9E-08	15.8	-2.87E-06	
79	20.37	0.89	26.92	Pyrazine, (1-methylethenyl)-	887	120	0.43	6.3E-06	18.6	-6.42E-06	Caramel-like, Chocolate, Nutty, Roasted
80	23.73	1.06	26.84	Unk30		161	0.6	3.4E-06	5	-2.11E-06	
81	24.23	0.63	26.68	2,2'-Bifuran	801	105	0.45	4.9E-06	13.5	-5.70E-05	
82	24.33	0.84	26.66	Unk31		91	0.53	2.5E-06	15.3	-2.23E-06	
83	36.8	1.21	26.59	Benzene, 1,1'-(2-methyl-1-propenylidene)bis-	815	193	S only	2.9E-06		1.61E-06	
84	19.6	1.57	26.43	Thiazole, 2,5-diethyl-4-methyl-	805	140	0.5	6.6E-06	19.1	-8.09E-07	Nutty, Green
85	28.63	0.82	26.08	Unk32		160	0.62	3.1E-06	3.3	-4.84E-06	
86	25.87	0.88	26.04	Unk33		72	0.6	5.3E-07	19.2	-7.61E-06	
87	18.43	0.51	25.98	1-Methoxyadamantane	841	109	0.45	4.5E-06	19.4	-3.59E-06	
88	41.6	0.56	25.91	Unk34		140	0.65	4.9E-06	13.5	-7.23E-06	
89	28.13	0.81	25.8	Unk35		115	0.57	2.3E-06	14.6	-1.57E-06	
90	19.13	1.31	25.79	2,3-Diethylpyrazine	885	133	0.02	5.3E-07	18.1	-1.49E-05	Nutty, Hazelnut
91	15.77	1.66	25.09	Benzene, 1,2,3-trimethyl-	896	136	0.54	2.3E-05	18.2	-9.10E-07	
92	13.73	1.44	25.05	Cyclohexene, 3-(1-methylethyl)-	803	81	0.46	5.9E-06	12.2	-3.01E-06	
93	21.43	1.04	24.65	2-Cyclopenten-1-one, 3,4,4-trimethyl-	811	133	0.57	3.4E-05	7.3	-2.60E-05	
94	22.27	1.22	24.59	Pyrazine, 2,5-dimethyl-3-(2-propenyl)-	823	148	0.46	5E-07	14.8	-1.81E-06	
95	19.77	1.54	24.58	2,5-Dimethyl-3-isopropylpyrazine	810	150	0.48	1.8E-05	15.1	-1.52E-06	
96	27.07	1.12	24.35	Furan, 2,2'-methylenebis[5-methyl-	862	176	0.42	1.7E-06	3.5	-2.32E-06	
97	13.07	0.83	24.15	trans-á-Ocimene	887	91	0.4	0.00012	19.7	-1.88E-05	
98	24.5	1.09	24.01	Unk36		131	0.53	0.00004	19.2	-3.07E-06	
99	25	0.19	23.72	2-Undecanone, 6,10-dimethyl-	923	71	1.36	0.00034	18.3	7.06E-06	Musty
100	23.37	1.22	23.55	2-Methyl-5,6,7,8-tetrahydroquinoxaline	820	148	0.45	8.5E-07	18.2	-8.07E-06	
101	15.03	1.62	23.51	Oxazole, 2,5-diethyl-4-methyl-	813	136	0.61	0.00037	13.8	-5.15E-06	

102	15.43	1.55	22.8	1,5,5-Trimethyl-6-methylene-cyclohexene	817	121	0.51	0.00078	9.6	-1.53E-06	
103	18.1	1.07	22.71	Unk37		123	0.39	0.00002	19.6	-4.68E-06	
104	29.33	0.08	22.59	Unk38		107	3.09	3.7E-05	17.3	7.40E-07	
105	20.63	1.71	22.53	3,6-Dimethyl-2,3,3a,4,5,7a-hexahydrobenzofuran	894	82	0.34	1.2E-06	19.7	-1.87E-06	Herbal
106	21.23	0.92	22.29	Unk39		108	0.24	1.1E-06	18.4	-1.71E-06	
107	19.87	0.99	22.21	Unk40		127	0.4	5.8E-06	19.8	-1.97E-06	
108	20.9	1.16	22.14	*n*-Pentylpyrazine	816	107	0.41	1.4E-05	19.9	-1.12E-05	
109	18.93	0.18	22.08	1,5-Hexadien-3-ol	809	57	0.56	2.2E-05	16.9	-4.44E-06	
110	14.2	1.92	22.06	Acetic acid, hexyl ester	908	61	0.62	1.3E-05	10.1	-3.99E-06	Fruity, Green, Sweet
111	23.63	1.43	22.06	Pyrazine, trimethyl-1-propenyl-, (*E*)-	805	162	0.52	9.5E-06	19	-2.37E-06	
112	28.17	0.66	22.02	Unk41		147	0.58	7.2E-05	5.1	-2.60E-06	
113	24.77	0.76	21.96	2-Acetyl-3-methylpyrazine	863	128	0.1	5.6E-06	8.1	-5.35E-05	Nutty, Hazelnut, Roasted
114	11	1.89	21.94	Acetic acid, pentyl ester	811	61	0.5	4.4E-05	13.5	-2.40E-06	Fruity, Banana, Pear, Apple
115	28.83	0.86	21.92	Unk42		174	0.58	7.7E-09	1.8	-4.66E-06	
116	20.37	1.92	21.84	Decanal	805	46	1.63	0.00013	10	5.85E-05	Waxy, Fatty, Aldehydic
117	25.73	0.65	21.8	3-Methyl-2-thiophenecarboxaldehyde	901	128	0.52	3.3E-05	11	-5.48E-05	Saffron, Camphoreous
118	22	1.01	21.73	Unk43		53	0.03	1.6E-06	12.4	-8.49E-06	
119	17.53	1.42	21.28	Pyridine, 2-(2-methylpropyl)-	880	120	C only	6.6E-06		-3.03E-06	Green, Bell Pepper
120	24.07	0.97	21.21	Pyrazine, 2-methyl-5-(1-propenyl)-, (*Z*)-	875	68	0.5	1.3E-05	19.9	-1.91E-04	
122	17.7	1.1	21.1	Pyrazine, trimethyl-	842	91	0.53	2.2E-05	19.7	-1.51E-06	
123	13.93	1.65	21.09	4,5-Dimethyl-2-isopropyloxazole	810	96	0.47	5.1E-05	12.9	-1.40E-05	
124	20.97	0.88	21.04	Unk44		154	0.55	1.1E-05	12.7	-4.01E-07	
125	15.97	1.83	21.03	Unk45		125	0.48	4.9E-05	11.7	-1.64E-05	
126	18.63	1.07	21.02	Benzene, 1-methoxy-3-methyl-	895	127	0.5	1.9E-05	7.8	-1.70E-06	Floral
127	27.97	0.9	20.97	Quinoline, 1,2,3,4-tetrahydro-	841	162	0.55	2.8E-06	11.9	-7.50E-06	Honey, Phenolic
128	24.5	0.79	20.97	Benzoxazole, 2-methyl-	846	137	0.56	2.8E-05	19.4	-1.09E-05	Tobacco, Burnt, Phenolic, Nutty
129	17.23	1.13	20.76	Pyrazine, 2-ethyl-6-methyl-	909	104	0.52	3.5E-05	18.7	-2.79E-06	Roasted, Hazelnut
130	20.37	1.52	20.73	Pyrazine, 3,5-diethyl-2-methyl-	809	149	0.43	0.00013	18.9	-7.43E-07	Nutty, Green

132	20.63	1.51	20.52	Pyrazine, 3,5-dimethyl-2-propyl-	866	120	0.46	3.2E-05	18.2	-2.04E-06	Nutty, Hazelnut, Burnt
133	22	0.66	20.35	2-Furancarboxaldehyde, 5-methyl-	922	140	0.53	3.3E-06	14.3	-9.10E-07	Spicy, Caramel-like, Maple, Coffee
134	19.9	1.66	20.32	Benzene, 1,2,4,5-tetramethyl-	856	150	0.56	1.9E-05	16.3	-1.40E-06	
135	36.47	1.38	20.31	Unk46		224	S only	0.00015		9.56E-07	
136	27.43	0.82	20.3	Unk47		95	0.54	8.7E-06	19.5	-2.78E-06	
137	21.63	1.01	20.25	1,6-Octadien-3-ol, 3,7-dimethyl-	889	85	0.72	2.6E-06	10	-3.72E-06	Citrus, Floral, Sweet
138	17.17	1.53	20.17	Phenol, 2,3,6-trimethyl-	811	136	0.46	3.3E-05	10.4	-4.05E-06	
139	19.67	1.35	20.09	Pyridine, 2-ethyl-4,6-dimethyl-	809	66	0.29	3E-06	17	-7.31E-06	
140	27.23	0.84	20.08	Unk48		140	0.43	0.00004	17.8	-7.73E-06	
141	19.27	1.32	20.01	Pyrazine, 2,5-diethyl-	889	57	0.25	3.4E-06	16.7	-1.66E-05	Nutty, Hazelnut
142	15.53	1.53	19.54	4-Penten-1-ol, propanoate	824	53	0.41	2.9E-06	18.7	-1.93E-06	
143	23.6	0.83	19.27	Unk49		70	0.46	7.1E-06	18.9	-9.04E-07	
144	14.07	1.54	19.23	Cyclohexane, 1-ethyl-2-methyl-, cis-	819	124	0.32	3.5E-06	13.2	-2.10E-06	
145	16.37	1.02	19.12	Pyridine, 2,6-dimethyl-	894	82	0.57	0.00019	16.1	-1.66E-06	Nutty, Bready, Cocoa
146	18.9	0.31	18.88	Acetic acid	854	57	0.19	0.00082	5.5	-7.88E-06	Sharp, Pungent, Sour
147	26.5	1.54	18.87	2,6-Octadien-1-ol, 3,7-dimethyl-, acetate	857	85	0.62	0.00048	11.4	-7.62E-07	Floral
148	19.37	1.28	18.84	Pyrazine, 2-methyl-6-propyl-	845	97	0.31	2.1E-06	19.1	-8.29E-05	Burnt, Hazelnut, Nutty
149	20.23	1.55	18.8	Unk50		84	2.7	3.3E-05	11.7	8.23E-06	
150	17.63	0.87	18.55	5-Ethylthiazole	904	86	0.48	8.5E-05	13.6	-9.77E-06	
151	34.1	0.86	18.16	Unk51		166	0.54	5.9E-05	18.6	-3.30E-06	
152	24.93	0.71	18.12	1-Methylenespiro[2.4]heptan-4-one	802	115	0.48	0.00002	17.2	-2.85E-06	
153	28.37	0.8	18.07	2-Naphthalenol	886	117	0.6	8.6E-07	10.8	-8.74E-04	
154	20.6	1.33	17.94	trans-3-Nonen-2-one	806	82	0.61	5.1E-06	15.2	-4.59E-06	
155	36	1.55	17.89	Propane, 2-cyclohexyl-2-phenyl-	801	119	S only	0.00023		7.51E-06	
156	16.7	1.23	17.88	Pyridine, 2-propyl-	875	93	0.5	9.9E-05	12.2	-9.17E-06	Green, Fatty, Roasted, Nutty
157	25.27	0.92	17.76	Benzaldehyde, 4-ethyl-	822	163	0.49	2.8E-05	19.1	-1.63E-05	Almond, Sweet, Anise, Cherry

158	14.7	1.6	17.7	2-Cyclopenten-1-one, 3-(dimethylamino)-2-methyl-	802	124	0.53	6.4E-05	17	-7.87E-07	
159	27.47	1.01	17.69	Unk52		176	0.58	6.8E-05	2.4	-1.86E-06	
160	22.8	0.74	17.67	1H-Pyrrole-2-carboxaldehyde, 1-ethyl-	839	105	0.54	4.1E-06	10.7	-2.79E-05	Burnt, Roasted
161	17.7	1.68	17.64	Unk53		107	0.52	0.00054	17.3	-1.43E-06	
162	24.17	0.7	17.55	Unk54		80	0.28	2.1E-06	14.8	-3.39E-06	
164	20.1	1.05	17.45	Unk55		78	0.45	0.0002	18.8	-1.44E-06	
165	16.83	0.91	17.4	Thiazole, 4,5-dimethyl-	926	46	0.46	4.3E-05	19.7	-9.24E-06	Roasted, Nutty, Green
166	14.57	1.4	17.34	Phenol, 2,5-dimethyl-	804	60	0.55	0.00022	15.4	-2.28E-06	Sweet, Phenolic, Smoky
167	23.57	0.69	17.33	5-Ethyl-2-furaldehyde	878	83	0.5	6.5E-05	11.7	-8.36E-06	
168	21.9	0.88	17.3	3,5-Octadien-2-one, (*E,E*)-	837	112	0.72	0.00013	17.6	-5.45E-06	Fruity, Green
169	16.1	1.17	17.25	2-Ethyl-4-methylthiazole	832	126	0.38	7.3E-05	19.3	-1.50E-06	Nutty, Coffee, Cocoa
171	18.47	1.51	16.99	Benzene, (1-methoxy-1-methylethyl)-	871	123	0.18	1.9E-05	19	-1.06E-05	
172	21.37	1.46	16.98	2,3-Dimethyl-5-n-propylpyrazine	809	85	0.5	3.5E-06	16.7	-1.23E-06	
173	23.57	0.92	16.82	5H-1-Pyrindine, 6,7-dihydro-	878	96	0.38	3.3E-09	4.8	-1.60E-05	
174	22.87	1.51	16.79	2-Isoamyl-6-methylpyrazine	837	52	C only	2.3E-05		-9.68E-07	
175	23.27	1.05	16.44	1,3-Cyclopentanedione, 2-isopentyl-	858	85	0.22	1.5E-05	14.1	-5.91E-06	
176	19.7	0.77	16.36	2-Acetyl-5-methylfuran	883	52	0.61	0.00039	11.6	-2.17E-06	Nutty, Coconut, Milky
177	16.3	0.87	16.32	Thiazole, 2-ethyl-	800	85	0.5	0.00012	18.8	-1.54E-05	Nutty, Green
178	35.03	1.04	16.31	Megastigmatrienone	831	89	0.6	0.00044	17.7	-1.27E-06	Swet, Nutty, Tobacco, Spicy
179	21.03	1.51	16.29	Pyrazine, 2-methoxy-3-(2-methylpropyl)-	890	124	0.57	0.00012	7.1	-8.85E-06	Green, Fresh
180	31.87	0.94	16.2	Unk56		135	0.57	6.8E-05	17.2	-9.36E-06	
181	20.93	0.96	16.1	2,4-Decadienal, (*E,E*)-	800	112	0.57	3.6E-05	11.2	-3.86E-06	Cucumber, Melon, Citrus, Nutty
182	22.33	1.35	16.09	2-Isoamylpyrazine	816	108	0.45	0.00017	18.1	-9.60E-06	
183	25.8	1.62	16.07	(*R*)-lavandulyl acetate	829	92	0.55	1.7E-05	16.5	-1.59E-06	Floral
184	25.5	0.91	15.96	Unk57		58	0.42	5.8E-08	11.7	-5.83E-06	
185	13.9	1.36	15.91	Oxepine, 2,7-dimethyl-	825	122	0.6	0.00061	15.9	-1.72E-06	
186	25.53	0.73	15.88	Bicyclo[4.1.0]heptan-2-one, 6-methyl-	800	152	0.33	3.9E-07	6.3	-2.71E-05	

187	32.5	0.78	15.81	Unk58		144	0.55	0.00014	9.7	-8.62E-06	
188	29.07	1.11	15.78	2,5,6-Trimethylbenzimidazole	808	189	0.44	8E-06	17.1	-4.79E-06	
190	29.07	0.79	15.77	Unk59		131	0.47	3.4E-05	17	-4.09E-06	
191	27.7	1.2	15.71	Unk60		93	0.61	3.5E-05	13.9	-7.37E-07	
193	15.43	1.26	15.68	Unk61		109	0.21	2.5E-05	17.9	-1.17E-06	
194	13.23	1.36	15.64	Unk62		69	0.55	0.00019	19.6	-7.45E-07	
195	28.53	0.96	15.4	2-Furanmethanethiol, 5-methyl-	821	96	0.35	6.2E-05	12.1	-5.18E-06	Coffee, Roasted
196	28.73	0.78	15.38	Unk63		163	0.58	0.00028	8.5	-2.82E-06	
197	22.1	0.99	15.36	Unk64		49	C only	2.8E-06		-2.91E-05	
198	32.03	0.82	15.03	Unk65		153	0.53	0.0001	16.6	-1.20E-06	
199	20.77	1.29	15	2-Ethyl-3,5-dimethylpyridine	839	134	0.52	0.00019	13.3	-3.89E-06	
200	31.63	0.78	14.92	2,7-Naphthalenediol	859	142	0.55	8.2E-05	15.1	-2.02E-04	
201	14.87	1.38	14.85	Pyridine, 2-ethyl-6-methyl-	914	93	0.55	0.00021	16.7	-7.01E-06	
202	14.97	1.99	14.82	5-Ethyl-2-isopropyl-4-methyloxazole	828	139	0.25	0.00019	18.9	-9.82E-07	
203	16.33	1.15	14.73	2-Isopropylpyrazine	844	85	0.35	0.00015	19.9	-1.61E-05	Nutty, Honey, Coffee
204	43.6	1	14.69	Unk66		119	8.07	0.00013		2.21E-06	
205	18.33	1.96	14.2	Benzene, 1-methyl-4-(1-methyl-2-propenyl)-	876	67	0.25	1.7E-08	20	-5.59E-07	
206	16.83	1.22	14.18	Unk67		123	0.41	1.2E-05	19.2	-1.11E-06	
207	17.13	0.73	14.05	Unk68		69	0.6	0.00016	15.1	-9.96E-07	
208	27.9	0.69	14.05	1H-Pyrrole, 1-(2-furanylmethyl)-	839	47	0.59	0.00042	15.5	-6.33E-06	Fruity, Coffee
209	20.73	1.48	14.04	2-Acetyl-3-ethylpyrazine	862	135	0.44	0.00031	8.1	-3.27E-06	Nutty, Cocoa
210	15.53	0.95	13.97	Pyrazine, 2,6-dimethyl-	915	83	0.56	8.8E-05	18.1	-1.29E-03	Cocoa, Roasted, Nutty
212	24.83	0.51	13.89	3-Thiophenecarboxaldehyde	817	109	0.58	3.9E-05	17.2	-3.80E-06	
213	16.57	0.89	13.86	Pyrazine, 2,3-dimethyl-	870	108	0.53	0.00064	13.7	-2.77E-06	Nutty, Cocoa, Coffee
214	29	1.09	13.69	Unk69		168	0.57	3.8E-05	18.3	-9.01E-07	
216	18.43	0.94	13.65	2-Cyclohexen-1-one, 3-methyl-	867	66	0.58	5.7E-05	16	-1.52E-06	Nutty, Caramel-like, Sweet
217	22.37	0.89	13.58	3-Acetyl-2,5-dimethyl furan	863	51	0.52	1.2E-06	18.8	-5.55E-05	Sweet, Nutty, Cocoa
218	18.43	0.83	13.46	Cyclopentanone, 3,4-bis(methylene)-	845	80	0.61	0.00043	12.5	-8.25E-06	
219	29.23	0.5	13.36	2-Furanacrylonitrile	801	119	0.55	0.00047	16.6	-4.58E-06	

220	33.37	0.8	13.34	5,7-Dimethylchromone-3-carboxaldehyde	801	146	0.5	0.00068	19	-2.02E-06	
221	31.9	0.73	13.26	Phenol, 4-ethyl-2-methoxy-	824	165	0.48	0.00061	9.5	-1.54E-05	Spicy, Smoky, Clove
222	31.7	1.02	13.21	(5-Oxo-2-thiophen-2-yl-cyclopent-1-enyl)acetic acid	801	150	0.65	8.9E-05	7.5	-5.63E-06	
223	29.63	0.61	13.18	2-Thiophenecarboxylic acid, 4-nitrophenyl ester	805	84	0.54	5.9E-05	13.5	-4.06E-05	
224	14.07	1.36	13.15	Unk70		110	C only	0.00015		-4.54E-07	
225	18.37	1.58	13.13	Unk71		81	C only	2.4E-05		-9.41E-07	
226	20.5	0.67	13.07	2-Cyclopenten-1-one, 3-methyl-	869	65	0.57	0.00082	17.8	-6.36E-06	Sweet, Fruity, Fatty
227	36.53	0.93	13.02	Unk72		118	0.42	0.00028	19.5	-9.52E-07	
228	14.43	1.62	13.01	2-Octanone	948	71	0.6	0.00014	18.2	-2.81E-06	Herbal, Earthy, Dairy
229	33.57	0.69	12.91	Unk73		160	0.51	0.00017	11	-3.68E-06	
230	24.33	1.71	12.9	Pyrazine, 2,5-dimethyl-3-(3-methylbutyl)-	815	109	0.62	0.00047	17.7	-6.06E-07	
231	23.17	0.99	12.87	5H-5-Methyl-6,7-dihydrocyclopentapyrazine	878	54	0.47	6.9E-05	19.6	-2.95E-05	Sweet, Nutty, Roasted, Coffee
233	18.63	1.37	12.84	Pyrazine, 2,6-diethyl-	882	52	0.3	4.3E-06	18.8	-2.84E-05	Nutty, Hazelnut
234	33.1	0.59	12.83	Unk74		124	0.53	0.00042	5.6	-7.00E-06	
235	11.6	1.62	12.81	Cyclohexanol, 1-methyl-4-(1-methylethenyl)-, acetate	916	80	0.47	0.00018	6.7	-4.32E-06	
236	22.63	0.47	12.79	Unk75		95	0.6	2.4E-05	17.5	-4.96E-06	
237	33.03	0.76	12.76	Unk76		103	0.54	0.00052	19.5	-3.43E-06	
238	15.07	0.98	12.75	4,4-Dimethyl-2-cyclopenten-1-one	836	67	0.65	0.00074	18.5	-9.25E-07	
239	32.8	0.83	12.75	Unk77		45	0.62	0.00034	19.5	-1.53E-06	
240	29.73	0.73	12.74	Unk78		106	0.06	8.9E-06	19.6	-8.97E-06	
241	33.93	0.63	12.64	Unk79		162	0.09	0.00014	13.1	-2.51E-06	
242	11.27	1.62	12.64	Heptanal	916	68	0.63	0.00092	17.6	-6.74E-06	Fresh, Fatty, Green, Herbal
243	31.43	0.92	12.56	8-Aminoquinaldine	801	64	0.56	0.00044	16.9	-7.50E-06	
244	2.7	0.66	12.54	Furan, 2,3-dihydro-	939	72	0.41	0.0005	17.3	-3.23E-05	
245	33.8	0.77	12.47	Thiophene, 2-phenyl-	890	157	0.29	3.2E-05	9.7	-3.88E-05	
247	27.6	0.61	12.44	3-Ethenyl-3-methylcyclopentanone	811	68	0.53	0.00068	3.8	-1.83E-05	

248	32.63	0.7	12.43	Phenol, 4-(1,1-dimethylpropyl)-	805	158	0.56	0.00037	11.8	-4.68E-06	
249	14.03	0.81	12.42	Pyrazine, methyl-	863	69	0.69	0.00088	17.3	-2.39E-06	
250	32.93	0.89	12.21	2-Benzothiazolamine, 6-methyl-	800	164	0.53	0.00057	8.6	-1.47E-05	
251	22.17	1.42	12.12	Pyridine, 2-pentyl-	907	78	0.53	9.7E-05	18.3	-1.63E-06	Fatty, Green, Herbal, Peanut-like
252	33.87	0.96	12.11	Acenaphthene	901	132	2.04	0.00003	2.8	4.22E-06	
253	13.5	1.33	12.09	2-Buten-1-ol, 3-methyl-, acetate	828	86	0.62	0.0002	15	-1.12E-06	Fruity, Sweet, Banana
254	33.37	0.92	11.98	Unk80		149	0.66	0.00034	10.5	-9.65E-07	
255	34.4	0.81	11.97	Unk81		75	0.48	0.0001	7.5	-4.10E-05	
256	11.17	0.84	11.97	Unk82		94	0.6	0.001	12.4	-5.64E-06	
257	27.77	1.36	11.97	2-Buten-1-one, 1-(2,6,6-trimethyl-1,3-cyclohexadien-1-yl)-, (*E*)-	901	135	0.73	0.00019	19.2	-1.98E-05	Sweet, Fruity, Rose
258	32.2	0.76	11.85	Unk83		71	0.57	0.00024	15.3	-7.20E-05	
259	18.43	1.26	11.83	2-Octenal, (*E*)-	903	71	0.67	0.00055	16.1	-8.51E-07	Fresh, Cucumber, Banana
260	7.77	1.52	11.82	Hexanal	880	52	0.61	0.00049	19.8	-1.72E-05	Fresh, Fatty, Fruity, Sweet
261	32.4	0.78	11.79	Unk84		153	0.63	0.00021	15.3	-4.34E-04	
262	29.27	0.87	11.75	Unk85		70	0.12	4.8E-05	14.5	-2.94E-06	
263	32.03	0.73	11.73	Benzene, 4-ethenyl-1,2-dimethoxy-	827	154	0.56	0.00035	11.1	-1.35E-05	Floral, Fruity, Green
264	33.7	0.85	11.67	Unk86		179	0.56	0.00075	18.2	-2.93E-06	
265	14.13	0.75	11.66	Butanoic acid, anhydride	874	48	0.59	0.00099	19.5	-1.59E-05	Butter
266	36.27	1	11.6	Cyclopenta[1,3]cyclopropa[1,2]cyclohepten-3(3aH)-one, 1,2,3b,6,7,8-hexahydro-6,6-dimethyl-	805	120	0.62	0.00038	14.9	-3.50E-06	
268	12.1	0.94	11.53	Pyridine, 2-methyl-	925	45	0.11	3.5E-05	9.3	-3.41E-05	Hazelnut, Nutty
269	14.73	1.77	11.52	Unk87		79	0.71	3.4E-06	13.2	-5.75E-07	
270	11.97	0.67	11.46	Pyridine	876	54	0.54	0.00082	15	-7.95E-06	Sour
271	22.17	0.68	11.32	1-Propanone, 1-(2-furanyl)-	874	82	0.59	0.00047	8.4	-4.00E-04	Fruity
272	36.9	0.77	11.23	1(2H)-Naphthalenone, 3,4-dihydro-6,7-dimethyl-	804	146	0.61	0.00097	6.1	-5.94E-06	
273	39.57	0.65	11.19	Unk88		103	0.64	0.00055	12.1	-1.39E-06	
274	36.4	0.8	11.18	Benzenamine, N,N-diethyl-2-methyl-	823	66	0.4	0.00042	17.5	-3.59E-05	
275	34.8	0.7	11.13	Unk89		131	0.61	0.00026	9.5	-1.49E-06	

276	37	0.57	11.1	Unk90		146	0.6	0.00034	11.1	-5.35E-07	
277	37.93	0.68	11.08	Quinoline, 2,7-dimethyl-	814	165	0.55	2.6E-05	7.8	-1.50E-06	
278	21.67	1.73	11.04	Pyrazine, 2,5-dimethyl-3-(2-methylpropyl)-	835	83	0.35	0.00035	15.4	-1.13E-05	
281	41	0.72	11.01	Unk91		227	0.59	0.00012	17.5	-1.25E-06	
282	44.3	0.91	10.96	Unk92		119	4.86	1.3E-05	13.1	5.14E-06	
283	31	1.13	10.79	Unk93		174	0.56	7.1E-05	14.6	-7.58E-07	
284	38.37	0.86	10.74	Unk94		64	0.53	0.00001	18.3	-1.01E-06	
285	34.17	0.71	10.69	Unk95		63	0.46	0.00063	15.8	-5.61E-06	
286	43.43	0.67	10.65	Unk96		146	0.57	0.00086	15.9	-3.33E-06	
287	12.63	1.52	10.6	2-Butenoic acid, 3-methyl-, ethyl ester	814	55	0.46	0.00021	15.8	-1.89E-06	
289	38.27	0.65	10.42	Unk97		106	0.66	3.8E-05	18.3	-1.13E-06	
292	38.83	1	10.37	Unk98		77	0.39	0.00031	5.3	-2.59E-06	
293	37.83	0.76	10.36	Unk99		94	0.64	2.7E-06	1	-6.23E-06	
294	37.53	0.77	10.33	Unk100		172	0.53	1.2E-05	6.1	-3.62E-06	
295	29.03	0.61	10.24	Unk101		94	0.39	0.00088	4.3	-2.08E-06	
296	38.6	0.76	10.18	Unk102		179	0.61	0.00035	13.7	-4.88E-06	
297	14.87	1.52	10.18	2-*n*-Butyl furan	860	56	0.46	6.7E-05	15.5	-8.80E-06	
298	40	0.81	10.15	Unk103		87	0.45	8.5E-05	17.7	-3.07E-06	
299	43.6	0.72	10.15	Unk104		81	0.24	8.9E-05	4	-5.47E-07	
300	38.6	0.43	10.15	3-Pyridinol	876	141	0.56	1.4E-05	0.8	-1.24E-06	
301	41.17	0.72	10.14	Unk105		131	0.64	4.1E-06	9.3	-2.67E-06	
302	17.87	0.72	10.04	4-Hydroxy-3-hexanone	823	57	0.6	0.00094	19	-1.48E-06	
303	25.67	0.53	10.02	1,4-Cyclohex-2-enedione	872	83	0.53	0.00065	14.8	-5.57E-06	
304	30.13	0.4	9.98	2-Thiophenemethanol	923	78	0.58	0.00082	18.7	-7.27E-06	Coffee, Roasted
305	28.33	0.76	9.95	Unk106		117	0.59	3.5E-05	12.9	-5.91E-06	
306	30.7	0.41	9.94	3-Acetylpyrrole	869	105	0.6	0.00063	10.1	-8.74E-04	Caramel-like, Sweet
307	38.5	0.75	9.9	Unk107		171	0.47	0.00011	8.5	-2.30E-06	
308	39.1	0.79	9.9	Unk108		185	0.59	0.00019	2.9	-5.02E-07	
309	34.63	0.83	9.84	Unk109		145	0.59	0.0004	16.8	-1.63E-06	
311	37.67	0.63	9.8	Unk110		162	0.6	0.00025	10.5	-4.57E-06	
312	41.23	0.5	9.74	2-Methyl-5-(1-butyn-1-yl)pyridine	922	117	0.46	0.0008	14	-3.21E-06	
313	28.3	0.51	9.72	2-Propenal, 3-(2-furanyl)-	898	124	0.62	0.00042	14	-2.28E-06	Spicy, Fruity, Vanilla, Nutty
314	27.77	1.02	9.7	2-Acetyl-3,4,6-trimethylpyrazine	801	55	0.2	2.7E-06	15.9	-1.89E-06	
315	38.77	0.68	9.66	Unk111		84	0.26	0.00034	13.1	-5.29E-07	
316	40.23	0.75	9.65	Unk112		117	0.12	0.00093	5.6	-8.83E-07	

317	34.07	0.58	9.63	Unk113		63	0.5	0.00024	17.5	-3.03E-06	
318	20.57	1.13	9.6	Acetic acid, phenyl ester	801	136	0.5	4.8E-05	12.4	-4.51E-06	Phenolic, Burnt
319	22.87	1.14	9.57	Unk114		81	0.43	4.9E-06	19.7	-1.33E-06	
320	41.8	0.79	9.54	Unk115		53	0.66	4.7E-05	16.9	-1.93E-06	
321	42.63	0.66	9.52	Unk116		61	0.57	0.00032	6.1	-9.27E-07	
322	42	0.66	9.48	Unk117		77	C only	6.5E-05		-1.43E-06	
323	40.5	0.77	9.46	Unk118		79	0.14	0.00045	11.2	-5.61E-07	
324	37.63	0.81	9.43	Unk119		90	0.35	8.2E-05	19.7	-2.75E-06	
325	35.03	1.3	9.39	1,1'-Biphenyl, 3,4-diethyl-	812	195	1.43	4.7E-05	16.9	3.11E-06	
326	35.87	0.61	9.35	à-Furfuryliden-à-furylmethylamine	802	98	0.66	0.00083	0.4	-7.11E-06	
327	40.1	0.52	9.34	Unk120		109	C only	0.00018		-1.18E-06	
329	20.5	0.81	9.3	Benzofuran	897	79	0.42	2.5E-06	16.3	-6.02E-06	Aromatic
330	39.17	0.41	9.29	Indole	903	70	0.58	0.0007	14	-2.04E-05	Floral
331	40.1	0.29	9.29	Unk121		75	0.71	0.00038	16.9	-1.78E-06	
332	32.73	0.69	9.27	2(1H)-Quinolinone, 4-methyl-	854	113	0.59	0.00059	13.2	-3.04E-05	
333	12.93	0.19	9.24	Furan, 2-pentyl-	935	97	0.56	0.00011	12.9	-1.52E-06	Fruity, Green
334	44.5	0.48	9.19	Unk122		118	0.25	0.00091	12.4	-2.03E-06	
335	38.03	0.72	9.07	Unk123		93	0.58	0.0003	13.4	-1.26E-06	
336	36.7	0.79	9.06	Unk124		145	0.56	0.00088	12.4	-5.23E-07	
337	41.9	0.52	9.05	Unk125		163	0.54	0.00047	13.1	-1.41E-06	
340	28.13	1.02	8.96	Unk126		110	0.59	9.3E-05	14.9	-1.75E-06	
341	9.83	0.84	8.96	Unk127		80	0.71	0.00047	4.5	-4.72E-06	
342	36.27	0.72	8.9	Unk128		161	0.55	0.00082	10.5	-9.21E-07	
343	41.8	1.7	8.86	Unk129		96	0.64	0.00092	17.5	-1.47E-06	
344	44.3	0.71	8.82	Unk130		115	0.31	0.00061	11.1	-6.58E-07	
345	39.3	0.48	8.77	Unk131		75	0.55	0.00012	16.8	-1.49E-06	
346	39.93	0.44	8.74	1H-Indole, 4-methyl-	905	87	0.56	0.00055	4.3	-8.62E-06	
347	39.4	0.38	8.69	Unk132		134	0.65	0.00056	9.9	-9.73E-07	
348	48.3	0.42	8.68	Unk133		180	0.71	0.0006	10.4	-1.42E-06	
349	40.97	0.26	8.64	Linoleic acid ethyl ester	836	100	C only	0.00022		-1.03E-05	Fatty, Fruity
351	42.73	0.54	8.57	Unk134		60	0.82	0.00051	12.5	-1.60E-06	
352	7.2	0.63	8.57	Unk135		81	0.6	0.00098	17.8	-2.26E-06	
353	42.13	0.69	8.53	[1,1'-Biphenyl]-4-carboxaldehyde	821	106	0.66	0.00002	4.8	-8.87E-07	
354	38.13	0.61	8.49	2-Furancarboxaldehyde, 5-[(5-methyl-2-furanyl)methyl]-	822	89	0.64	0.00059	15.8	-4.15E-06	
355	34.23	0.5	8.47	Benzenamine, 4-methoxy-N-methyl-	806	104	0.55	0.00065	13.3	-1.22E-06	

356	42.1	0.61	8.43	3-Ethyl-5,6,7,8-tetrahydroquinoline	828	93	0.31	0.001	6.8	-3.76E-06	
357	43.27	0.5	8.4	Unk136		159	0.64	0.00075	18.3	-9.65E-07	
358	29.53	0.51	8.35	Phenylethyl Alcohol	905	110	0.57	0.00071	17	-6.60E-07	Floral, Sweet, Honey, Bready
359	43.13	0.55	8.34	Unk137		94	0.24	0.00055	19.4	-4.05E-06	
360	34.33	0.61	8.32	1H-1,3a-Ethanopentalen-5(4H)-one, 2,3-dihydro-	801	137	C only	0.00022		-1.53E-05	
361	43.87	0.37	8.31	3-Hydroxy-2-methylbenzaldehyde	833	75	0.74	0.00072	17.3	-1.75E-06	
362	31.57	0.92	8.3	Unk138		104	0.25	9.9E-07	2.8	-8.90E-06	
364	45.07	0.57	8.22	1,4-Benzenediol, 2,6-bis(1,1-dimethylethyl)-	802	222	1.38	0.00086	4.2	2.41E-06	
365	14.27	1.02	8.21	Thiazole, 2,4-dimethyl-	892	45	0.52	0.0005	5.8	-6.65E-06	Coffee, Tea, Roasted
366	12.77	1.32	8.13	Unk139		112	C only	0.00095		-5.64E-06	
370	43	0.55	8.05	Unk140		125	3.91	0.00013	10.4	-8.00E-07	
371	37.4	1.06	7.98	Unk141		160	0.31	0.00088	14.2	-4.02E-07	
373	34.53	0.49	7.91	Nonanoic acid	889	120	0.64	0.00088	17.4	-5.10E-06	Waxy, Dairy, Fatty
375	33.97	0.34	7.79	1H-Pyrrole-2-carboxaldehyde, 1-methyl-	901	53	0.59	0.00088	5.6	-2.76E-05	
376	43	0.38	7.76	Unk142		91	0.2	0.00015	6.4	-1.52E-06	
377	41.37	0.63	7.72	Unk143		131	C only	1.6E-06		-2.71E-07	
379	43.53	0.42	7.59	Unk144		94	0.72	0.00074	12.2	-6.39E-07	
380	25	1.41	7.52	(1-Methylpenta-1,3-dienyl)benzene	821	145	0.48	1.5E-06	15.5	-1.19E-06	
383	30.4	0.46	7.32	Maltol	939	81	0.55	0.00035	15.5	-9.18E-06	Sweet, Caramel, Candy, Baked
392	34.93	0.51	7.07	5-Acetoxymethyl-2-furaldehyde	918	53	0.62	0.00089	18.7	-3.21E-06	Baked, Bread
395	13.2	0.59	7.01	Thiazole	863	59	0.6	0.00036	13.9	-9.48E-07	Nutty
398	29.9	0.61	6.92	Carbonic acid, ethyl phenyl ester	827	94	0.66	0.0005	6.3	-1.06E-05	
401	33.77	0.41	6.9	Unk145		94	0.6	0.00051	7.5	-3.48E-06	
408	15.33	0.98	6.58	Pyrazine, 2,5-dimethyl-	907	72	0.53	0.00056	19.2	-6.26E-06	Cocoa, Roasted, Nutty
491	25.4	0.81	0.14	2-Butanone, 4-(5-methyl-2-furanyl)-	901	120	0.42	2.5E-06	10.6	-2.01E-07	Bitter, Roasted

[2] G.S. Ochoa, S.E. Prebihalo, B.C. Reaser, L.C. Marney, R.E. Synovec, Statistical inference of mass channel purity from Fisher ratio analysis using comprehensive two-dimensional gas chromatography with time of flight mass spectrometry data, J. Chromatogr. A 1627 (2020) 461401. https://doi.org/10.1016/j.chroma.2020.461401.

[3] The Good Scents Company, The Good Scents Company Information System, (2018). http://www.thegoodscentscompany.com/.

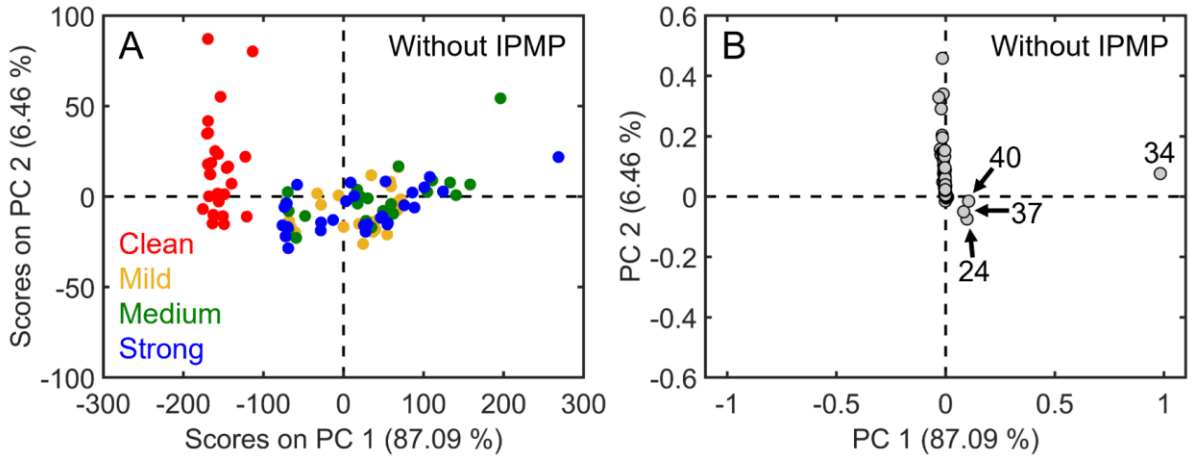


Figure B.3. Results from the PCA model constructed using the normalized intensity measured at the S-ratio *m/z* for all the discovered hits except for IPMP. (A) Scores plot for the model. Each class is colored accordingly: clean (red), mild PTD (yellow), medium PTD (green), and strong PTD (blue). (B) Loadings plot for the model shown in (A). Four highly loaded hits (hit #24, 34, 37, and 40) on PC 1 are labeled.

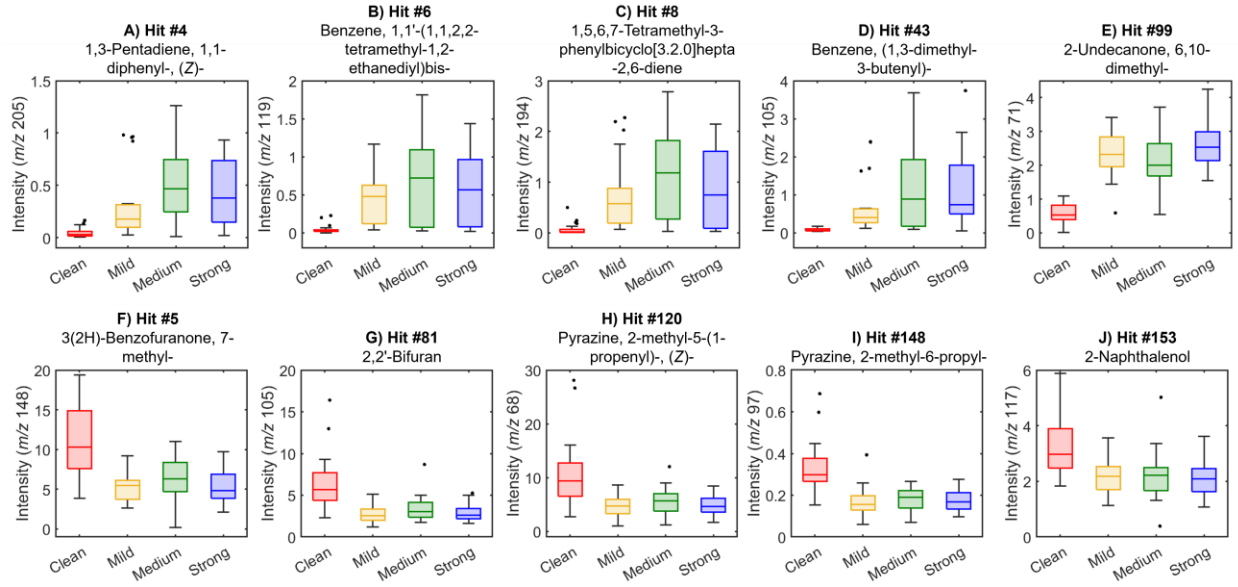


Figure B.4. Additional box-and-whiskers plots relating the intensity measured at the S-ratio *m/z* to their PTD odor attribution for analytes, which were highly loaded in the PLS model. The top row highlights analytes with signals larger in the PTD affected samples: (A) 1,3-pentadiene, 1,1-diphenyl-, (*Z*)-, (B) benzene, 1,1'-(1,1,2,2-tetramethyl-1,2-ethanediyl)bis-, (C) 1,5,6,7-tetramethyl-3-phenylbicyclo[3.2.0]hepta-2,6-diene, (D) benzene, (1,3-dimethyl-3-butenyl)-, and (E) 2-undecanone, 6,10-dimethyl-. The bottom row highlights analytes with signals larger in the clean coffee samples: (F) 3(2H)-benzofuranone, 7-methyl-, (G) 2,2'-bifuran, (H) pyrazine, 2-methyl-5-(1-propenyl)-, (*Z*)-, (I) pyrazine, 2-methyl-6-propyl-, and (J) 2-naphthalenol**.**

Appendix C

This appendix is reproduced from the Electronic Supplementary Material of C. N. Cain*, P.E. Sudol*, K. L. Berrier*, R. E. Synovec, Development of Variance Rank Initiated-Unsupervised Sample Indexing for Gas Chromatography-Mass Spectrometry Analysis, Talanta 233 (2021), 122495.
* These authors contributed equally.

Table C.1. Simulation parameters.

Parameter	Value
Total separation time, t_{sep}	50 s
Number of analytes, m	50
Peak capacity, n_c	50
Saturation factor, α	1
Peak width-at-base, w_b	1 s
Data collection rate	10 Hz
Number of mass channels, n	360
Number of chromatographic replicates	10 per class, 20 total
Number of simulations	100
Average peak height in TIC, prior to changing concentrations of 4 analytes	103,000 ± 14,000
Average standard deviation of the noise, per *m/z*	90 ± 12
Average signal-to-noise ratio in TIC, *S/N*	20
Within class variation, *%RSD*	30%

Table C.2. List of yeast samples analyzed. Sample names are labeled in the following order: culture (A, B, C), extraction replicate (1, 2, 3), class (R, DR): injection replicate (1, 2, 3, 4).

Repressed (R)	Derepressed (DR)
A1R:1	A1DR:2
A1R:3	A1DR:3
B1R:2	B1DR:1
B1R:3	B1DR:2
C1R:1	C1DR:2
C1R:2	C1DR:3

Experimental Conditions for Yeast Metabolome Data set

The data set consists of two classes of yeast, one of which was provided with glucose to enact fermentation (repressed, R) and the other was provided with ethanol to cause respiration (derepressed, DR). Three yeast cultures for each class were maintained (A, B, C), followed by three extractions of each culture, and four injection replicates. Metabolites were extracted with ethanol and then trimethylsilyl (TMS) derivatized. The yeast extracts were analyzed using an Agilent 6890N gas chromatograph (GC) equipped with an Agilent 7693 auto-injector (Agilent Technologies, USA) coupled to a LECO Pegasus III TOFMS with a 4D thermal modulator upgrade (LECO, USA). A sample volume of 1 μL was injected onto the first dimension ^{1}D column (RTX-5MS, 20 m × 250 μm i.d. × 0.5 μm, Restek, USA), which was initially held at 60 °C for 0.25 min and then increased at 8 °C/min to 280 °C and held for 10 min. The ^{1}D column effluent was trapped, refocused, and reinjected onto the second dimension ^{2}D column (RTX-200MS, 2 m × 180 μm i.d. × 0.2 μm, Restek, USA) with a sampling density of 1.5 s. The ^{2}D column was set to be 10 °C offset from the temperature of the ^{1}D column. Data were collected at a rate of 100 spectra/s (100 Hz) following a 5 min solvent delay. Further details can be found in the original reports [1,2].

[1] R.E. Mohler, K.M. Dombek, J.C. Hoggard, E.T. Young, R.E. Synovec, Comprehensive two-dimensional gas chromatography time-of-flight mass spectrometry analysis of metabolites in fermenting and respiring yeast cells, Anal. Chem. 78 (2006) 2700–2709. https://doi.org/10.1021/ac052106o.

[2] R.E. Mohler, K.M. Dombek, J.C. Hoggard, K.M. Pierce, E.T. Young, R.E. Synovec, Comprehensive analysis of yeast metabolite GC×GC-TOFMS data: Combining discovery-mode and deconvolution chemometric software, Analyst. 132 (2007) 756–767. https://doi.org/10.1039/b700061h.

Experimental Conditions for Human Cancer Data set

The data set consists of two classes of human saliva, one of which was from control patients and the other was collected from patients with head and neck cancer. The collected saliva samples (1 mL) were placed into a vial along with HCl (125 μL, 5 M) and NaCl (100 mg). The prepared samples were then extracted using a carboxen/polydimethylsiloxane (CAR/PDMS, 75 μm, Supelco, USA) solid-phase microextraction (SPME) fiber at 38 ºC for 45 min. The SPME fiber was desorbed in inlet of an Agilent 7890B GC (Agilent Technologies, USA) at 250 ºC for 6 min. The headspace volatiles were separated on a BP-20 (60 m × 250 μm, × 0.25 μm) capillary column using the following temperature program: 45 ºC for 5 min, increased at 2 ºC/min to 150 ºC, held at 150 ºC for 10 min, ramped up to 220 ºC at 15 ºC/min, and then held at 220 ºC for 15 min. After a 5 min solvent delay, the Agilent 5977A quadrupole mass selective detector (Agilent Technologies, USA) began data acquisition to collect mass channels 30 – 300 *m/z* using an ionization energy of 70 eV. Metabolites with a match score greater than 75 % and were found in over half of the samples were selected for further analysis. Additional details can be found in the original report [3].

[3] R. Taware, K. Taunk, J.A.M. Pereira, A. Shirolkar, D. Soneji, J.S. Câmara, H.A. Nagarajaram, S. Rapole, Volatilomic insight of head and neck cancer via the effects observed on saliva metabolites, Sci. Rep. 8 (2018) 17725. https://doi.org/10.1038/s41598-018-35854-x.

Table C.3. Entire VRI-USI hit list, ranked by RSD^2, for the simulated data set containing a background variance of 0.09. Sample index assignments are shown for $k = 2$. Hits shaded in green had matching sample index assignments and were correctly clustered into the two simulated classes. The concentration ratio, [Class A]/[Class B], and *p*-value obtained from a *t*-test is also provided.

Hit Number	t_R (s)	RSD^2	Sample Index Assignments	[Class A]/[Class B]	*p*-value
1	25.34	0.38	Samples 1-10; Samples 11-20	0.33	< 0.001
2	15.51	0.37	Samples 1-10; Samples 11-20	2.99	< 0.001
3	5.88	0.22	Samples 1-10; Samples 11-20	0.50	< 0.001
4	16.96	0.21	Samples 1-10; Samples 11-20	2.00	0.001
5	33.82	0.13	Samples 1-10; Samples 11-20	1.51	0.009
6	29.00	0.13	Samples 1-10; Samples 11-20	0.67	0.009
7	32.08	0.10	Samples 1,4,6-10; Samples 2,3,4,11-20	1.00	0.993
8	41.43	0.09	Samples 1-7; Samples 8-20	1.00	0.985
9	11.18	0.09	Samples 1-5,14,16-18,20; Samples 6-13,15,19	1.00	0.995
10	24.28	0.09	Samples 1,2,7-10,12,16,17,19,20; Samples 3-6,11,13-15,18	0.99	0.960
11	9.15	0.09	Samples 1,4,8,11-16,19,20; Samples 2,3,5-7,9,10,17,18	0.99	0.968
12	35.74	0.09	Samples 1-3,7,8,11-13,15,20; Samples 4-6,9,10,14,16-19	1.00	0.996
13	26.40	0.09	Samples 1-4,9-11,13,14,19,20; Samples 5-8,12,15-18	1.00	0.998
14	19.75	0.09	Samples 1,6,8,15,19; Samples 2-5,7,9-14,16-18,20	1.00	0.998
15	4.43	0.09	Samples 1,2,4,5,7,8,12,14-17,19; Samples 3,6,9-11,13,18,20	1.01	0.921
16	13.20	0.09	Samples 1,4,6,8,11-14,19; Samples 2,3,5,7,9,10,15-18,20	1.00	0.988
17	2.50	0.09	Samples 1-4,6-8,10-13,16,18-20; Samples 5,9,14,15,17	1.00	0.986
18	43.06	0.09	Samples 1-3,5,7-9,12-16,19; Samples 4,6,10,11,17,18,20	1.00	0.972
19	7.03	0.09	Samples 1-4,7,9-14,17,20; Samples 5,6,8,15,16,18,19	1.00	0.993
20	12.33	0.09	Samples 1-4,6,8,10-13,15-17,19; Samples 5,7,9,14,18,20	0.99	0.964
21	30.64	0.09	Samples 1,3-6,8,9,12,13,15,16,19,20; Samples 2,7,10,11,14,17,18	0.99	0.949
22	40.17	0.09	Samples 1,2,4,5,9,11,14,16,18; Samples 3,6-8,10,12,13,15,17,19,20	1.01	0.916
23	22.74	0.09	Samples 1,2,5-7,10,12,15-18,20; Samples 3,8,9,11,13,14,19	1.00	0.981
24	18.21	0.09	Samples 1,2,4,5,8-12,14,15,19,20; Samples 3,6,7,13,16-18	1.00	0.988
25	44.41	0.09	Samples 1,6,10,13-16; Samples 2-5,7-9,11,12,17-20	1.00	0.987
26	20.91	0.09	Samples 1,2,6,8,11,15,16,20; Samples 3-5,7,9,10,12-14,17-19	0.99	0.953

Table C.4. Entire VRI-USI hit list, ranked by RSD^2, for the simulated data set containing a background variance of 0.09. Sample index assignments are shown for $k = 3$. It was determined that none of the hits had matching sample index assignments.

Hit Number	t_R (s)	RSD^2	Sample Index Assignments
1	25.34	0.38	Samples 1-10,19; Samples 11,13,15,16; Samples 12,14,17,18,20
2	15.51	0.37	Samples 1,6-8,10; Samples 2,3,11-20; Samples 4,5,9
3	5.88	0.22	Samples 1-7,9,10,12,14; Samples 8,13,17,18; Samples 11,15,16,19,20
4	16.96	0.21	Samples 1,3-5,9,10; Samples 2,7,12,17,19; Samples 6,8,11,13-16,18,20
5	33.82	0.13	Samples 1,2,6,8,9; Samples 3,5,7,16,18,19; Samples 4,10-15,17,20
6	29.00	0.13	Samples 1,4-9,11,18; Samples 2,3,10,12,13,19; Samples 14-17,20
7	32.08	0.10	Samples 1,5,8,14,15,19,20; Samples 2-4,7,10-13,18; Samples 6,9,16,17
8	41.43	0.09	Samples 1,8,13,15,17; Samples 2-4,7,9-12,16,19,20; Samples 5,6,14,18
9	11.18	0.09	Samples 1,7,9,10,14; Samples 2,3,6,19,20; Samples 4,5,8,11-13,15-18
10	24.28	0.09	Samples 1,5,17; Samples 2,4,7,9,12,15,16; Samples 3,6,8,10,11,13,14,18-20
11	9.15	0.09	Samples 1,4,9,10,13-15,18,19; Samples 2,5-8,11,12,16,20; Samples 3,17
12	35.74	0.09	Samples 1,2,11-13; Samples 3,5,6,8,15,17-20; Samples 4,7,9,10,14,16
13	26.40	0.09	Samples 1-3,5,6,9,10,12,14,19,20; Samples 4,11,13; Samples 7,8,15-18
14	19.75	0.09	Samples 1,3,4,7,9,13,14,16,18; Samples 2,5,10,11,12,17,20; Samples 6,8,15,19
15	4.43	0.09	Samples 1,2,12,15,19; Samples 3,6,9,10,11,13,18,20; Samples 4,5,7,8,14,16,17
16	13.20	0.09	Samples 1,4,11; Samples 2,6,8-10,12-16,19; Samples 3,5,7,17,18,20
17	2.50	0.09	Samples 1-3,6-8,11,12,16,18,19; Samples 4,10,13,20; Samples 5,9,14,15,17
18	43.06	0.09	Samples 1,2,5,9,12-14,16,19; Samples 3,7,8,15; Samples 4,6,10,11,17,18,20
19	7.03	0.09	Samples 1,3,9,14,20; Samples 2,4,7,10-13,16,17; Samples 5,6,8,15,18,19
20	12.33	0.09	Samples 1-3,8,10,11,13,16,17; Samples 4,6,12,15,19; Samples 5,7,9,14,18,20
21	30.64	0.09	Samples 1,3,5,6,8,9,12,13,15,16; Samples 2,7,10,11,14,17,18; Samples 4,19,20
22	40.17	0.09	Samples 1,2,4,5,11,14,16,18; Samples 3,6,8,10,12,15,17,19; Samples 7,9,13,20
23	22.74	0.09	Samples 1,6,10,13,15,17,18,20; Samples 2,5,7,12,16; Samples 3,4,8,9,11,14,19
24	18.21	0.09	Samples 1,4,5,8,14,15; Samples 2,9-12,19,20; Samples 3,6,7,13,16-18
25	44.41	0.09	Samples 1,6,10,13-16; Samples 2,12,17; Samples 3-5,7-9,11,18-20
26	20.91	0.09	Samples 1,6,15,16,20; Samples 2,5,8-11,13,17; Samples 3,4,7,12,14,18,19

Table C.5. Entire VRI-USI hit list, ranked by RSD^2, for the simulated data set containing a background variance of 0.09 – 0.25. Sample index assignments are shown for $k = 2$. Hits shaded in green had matching sample index assignments and were correctly clustered into the two simulated classes. The concentration ratio, [Class A]/[Class B], and *p*-value obtained from a *t*-test is also provided.

Hit Number	t_R (s)	RSD^2	Sample Index Assignments	[Class A]/[Class B]	*p*-value
1	15.51	0.56	Samples 1-10; Samples 11-20	2.96	< 0.001
2	25.34	0.40	Samples 1-10; Samples 11-20	0.34	< 0.001
3	5.88	0.33	Samples 1-10; Samples 11-20	0.50	< 0.001
4	16.96	0.28	Samples 1-10; Samples 11-20	2.01	0.001
5	12.33	0.24	Samples 1-7,9,10,14; Samples 8,11-13,15-20	1.00	0.998
6	7.03	0.24	Samples 1,5,6,8,9,11,14,16-19; Samples 2-4,7,10,12,13,15,20	1.00	0.994
7	33.82	0.23	Samples 1-10; Samples 11-20	1.50	0.045
8	32.08	0.22	Samples 1-3,9-12,15,19; Samples 4-8,13,14,16-18,20	1.00	0.997
9	30.64	0.22	Samples 1-9,11,20; Samples 10,12-19	1.00	0.998
10	40.17	0.21	Samples 1-4,8,9,11-13,15; Samples 5-7,10,14,16-20	1.00	0.998
11	4.43	0.20	Samples 1,5,6,9-11,13,15,16; Samples 2-4,7,8,12,14,17-20	1.00	0.992
12	29.00	0.19	Samples 1-10; Samples 11-20	0.67	0.032
13	29.38	0.19	Samples 1,2,4,7,10,11,13,16-20; Samples 3,5,6,8,9,12,14,15	0.67	0.032
14	36.22	0.19	Samples 1,4,7,8,12-14,17,19; Samples 2,3,5,6,9-11,15,16,18,20	1.00	0.999
15	35.74	0.19	Samples 1,3-12,14-17; Samples 2,13,18-20	1.01	0.962
16	42.20	0.18	Samples 1,2,6,7,10,11,13,15,19,20; Samples 3-5,8,9,12,14,16-18	1.00	0.978
17	22.74	0.16	Samples 1-4,6-9,11,13,14,16,18-20; Samples 5,10,12,15,17	1.00	0.998
18	20.91	0.15	Samples 1-4,9,10,13,14,16-18; Samples 5-8,11,12,15,19,20	1.00	0.984
19	43.83	0.15	Samples 1,2,4-8,11,13,14,16,18,19; Samples 3,9,10,12,15,17,20	1.00	0.992
20	18.21	0.14	Samples 1,4,5,9-11,13,14,16,19; Samples 2,3,6-8,12,15,20	1.00	0.989
21	9.15	0.13	Samples 1,4,6,8,10,12,13,19,20; Samples 2,3,5,7,9,11,14-18	1.00	0.988
22	26.40	0.11	Samples 1,2,4,5,8,11,12,14,15,20; Samples 3,6,7,9,10,13,17-19	1.00	0.986
23	13.20	0.10	Samples 1-4,9,10,12,14,16-20; Samples 5-8,11,13,15	1.00	0.993
24	11.18	0.10	Samples 1,2,7,9-11,13,15-17,19; Samples 3-6,8,12,14,18,20	1.00	0.994
25	2.50	0.09	Samples 1,2,7,10-12,17-19; Samples 3-6,8,9,13-16,20	1.00	0.976

Table C.6. Entire VRI-USI hit list, ranked by RSD^2, for the simulated data set containing a background variance of 0.09 – 0.25. Sample index assignments are shown for $k = 3$. It was determined that none of the hits had matching sample index assignments.

Hit Number	t_R (s)	RSD^2	Sample Index Assignments
1	15.51	0.56	Samples 1,10,11,13,15-18,20; Samples 2,4-9; Samples 3,12,14,19
2	25.34	0.40	Samples 1,2,5-9; Samples 2,4,6,11,13,14,19; Samples 3,4,10,12,14,16,19
3	5.88	0.33	Samples 1,3,5,7-10,18,20; Samples 2,4,6,11,13,14,19; Samples 12,15-17
4	16.96	0.28	Samples 1-3,5,9-11,14; Samples 4,7,8; Samples 6,12,13,15-20
5	12.33	0.24	Samples 1,4,8,11,13,15,16,18; Samples 2,5,6,17; Samples 3,7,9,10,12,14,19,20
6	7.03	0.24	Samples 1,6,10,15,17; Samples 2,7-9,13,14,16,19,20; Samples 3-5,11,12,18
7	33.82	0.23	Samples 1,4,5,12,20; Samples 2,7,8,13,15,18; Samples 3,6,9-11,14,16,17,19
8	32.08	0.22	Samples 1-3,7,11,13; Samples 4,6,10,14,16-18,20; Samples 5,8,9,12,15,19
9	30.64	0.22	Samples 1,6,8,13-15,18; Samples 2,3,7,10,12,16,17,19; Samples 4,5,9,11,20
10	40.17	0.21	Samples 1,5,8,9,12,15,18-20; Samples 2,4,6,10,13,14,16; Samples 3,7,11,17
11	4.43	0.20	Samples 1,5,6,10,13,19,20; Samples 2,12,16; Samples 3,4,7-9,11,14,15,17,18
12	29.00	0.19	Samples 1,9,11,13,15,19; Samples 2,4,5,8,16,17; Samples 3,6,7,10,12,14,18,20
13	29.38	0.19	Samples 1,4,7,8,12-14,17,19; Samples 2,5,6,16,20; Samples 3,9-11,15,18
14	36.22	0.19	Samples 1,3,6,7,10,11,16,17; Samples 2,4,5,8,9,12,14,15; Samples 13,18-20
15	35.74	0.19	Samples 1,2,4,7,13,17,20; Samples 3,5,6,8,9,12,14,15; Samples 10,11,16,18,19
16	42.20	0.18	Samples 1,2,6,7,10,11,13,15,20; Samples 3-5,16-19; Samples 8,9,12,14
17	22.74	0.16	Samples 1,3,6,7,9,18-20; Samples 2,4,8,11,13,14,16; Samples 5,10,12,15,17
18	20.91	0.15	Samples 1,2,4,7-10,12,14,16,18,19; Samples 3,13,17; Samples 5,6,11,15,20
19	43.83	0.15	Samples 1,2,7,8,12,14,16; Samples 3,9,10,15,17,20; Samples 4-6,11,13,18,19
20	18.21	0.14	Samples 1,10,17,18; Samples 2,3,6-8,12,15,20; Samples 4,5,9,11,13,14,16,19
21	9.15	0.13	Samples 1,4,6,8,10,12,13,19,20; Samples 2,5,11,14,15,18; Samples 3,7,9,16,17
22	26.40	0.11	Samples 1,2,4,5,11,16,20; Samples 3,6,7,9,10,13,17,19; Samples 8,12,14,15,18
23	13.20	0.10	Samples 1-4,9,10,14,16-20; Samples 5,8,13,15; Samples 6,7,11,12
24	11.18	0.10	Samples 1,2,7,9,10,13,15,17; Samples 3,11,12,16,19; Samples 4-6,8,14,18,20
25	2.50	0.09	Samples 1,2,10-12,17,19; Samples 3-5,7,8,14,15,18,20; Samples 6,9,13,16

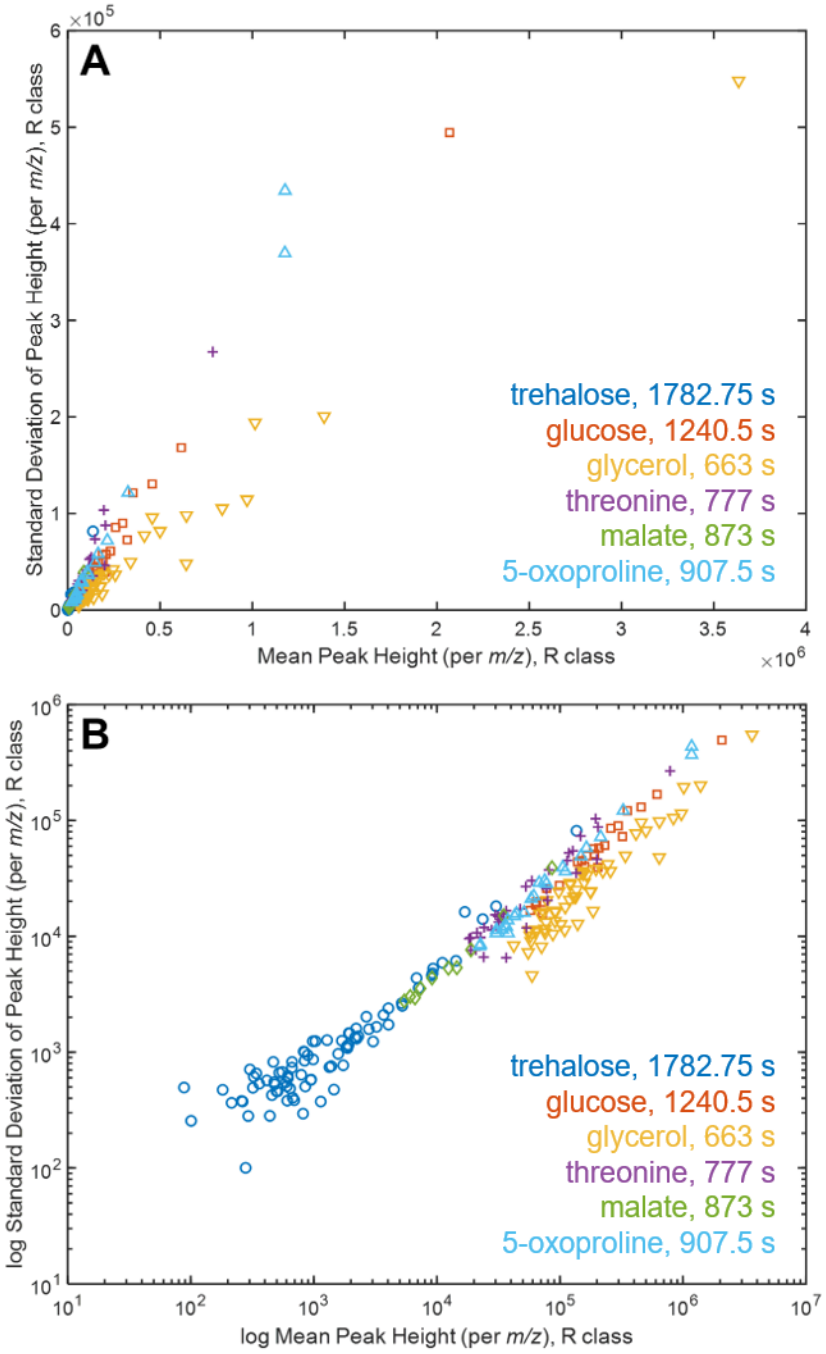


Figure C.1. (A) Scatter plot of the standard deviation versus mean of the peak height of *m/z* (with at least 5 samples that passed the threshold) for six analytes in the repressed samples individually: trehalose (blue circle), glucose (orange square), glycerol (yellow upside down triangle), threonine (purple plus sign), malate (green diamond), and 5-oxoproline (light blue triangle). (B) Logarithmically transformed standard deviation and peak height data from (A).

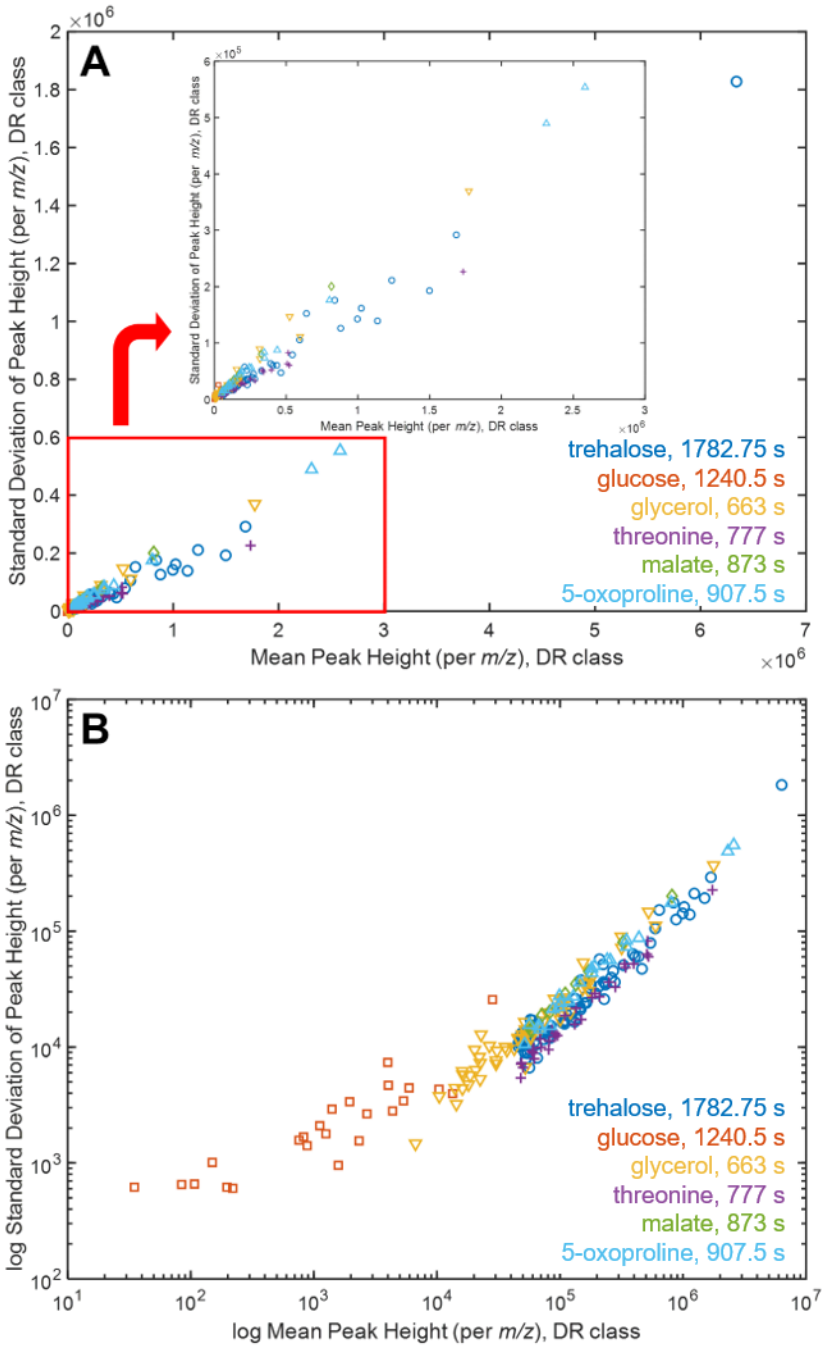


Figure C.2. (A) Scatter plot of the standard deviation versus mean of the peak height of *m/z* (with at least 5 samples that passed the threshold) for six analytes in the derepressed samples individually: trehalose (blue circle), glucose (orange square), glycerol (yellow upside down triangle), threonine (purple plus sign), malate (green diamond), and 5-oxoproline (light blue triangle). Zoom in from 0 to 3×10^6 in peak height provided inset. (B) Logarithmically transformed standard deviation and peak height data from (A).

Table C.7. List of *k*-means clustering results for the yeast metabolome data set using $k = 3$. Analytes shaded in the same colors have matching sample index assignments.[a-f]

Hit Number	Analyte	t_R (s)	RSD^2	Sample Index Assignments
1	Unk1	956.25	1.83	Samples 1,3,5-7,9; Samples 2,4,8,10,11; Sample 12
2	Unk2	1220.625	1.63	Samples 1,3,5-7,9,11; Samples 2,4,8,10; Sample 12
3	Unk3	1503	1.47	Sample 1,3,6,7,11; Samples 2,4,5,8,12; Samples 9,10
4	Glucopyranose	1282.5	1.38	Samples 1-3,6; Samples 4,5; Samples 7-12
5	Glucose	1240.5	1.18	Samples 1,3,6,7,9-11; Samples 2,4,5,8; Sample 12
6	Glucose	1227	1.17	Samples 1,2,5,6; Samples 3,4,7,9,12; Samples 8,10,11
7	Trehalose	1782.75	1.11	Samples 1,3,6,7,9; Samples 2,4,5,8,10; Samples 11,12
8	Unk4	699.75	0.836	Samples 1-3,5-7,9,10; Samples 4,8,12; Sample 11
9	Malate	873	0.809	Samples 1,5-7,9,10; Samples 2,4,12; Samples 3,8,11
10	5'-*S*-Methyl-5'-thioadenosine	1804.5	0.799	Samples 1,2,4-6; Samples 3,7-11; Sample 12
11	Unk5	924	0.687	Samples 1,2,4-6; Samples 3,7-11; Sample 12
12	Homoserine	835.5	0.612	Samples 1,3,6,7,9; Samples 2,8,10-12; Samples 4,5
13	Citrate	1160.25	0.562	Samples 1,3,6,7,9,10; Samples 2,4,5,8,11; Sample 12
14	Tyrosine	1247.25	0.537	Samples 1,3,6,7,9,10; Samples 2,4,5,8,11; Sample 12
15	Ornithine	1156.5	0.495	Samples 1,2,8,9,12; Samples 3-5,10,11; Samples 6,7
16	Lysine	1232.25	0.439	Samples 1-6; Samples 7-11; Sample 12
17	Unk6	1170.75	0.406	Samples 1,3,6,7,9,10; Samples 2,4,5,8,11; Sample 12
18	Unk7	696	0.404	Samples 1-6; Samples 7,9,10; Samples 8,11,12
19	Glucopyranose	1216.5	0.404	Samples 1,3,9; Samples 2,8,12; Samples 4-7,10,11
20	Unk8	1323	0.384	Samples 1,7,9,10; Samples 2,4,5,8,12; Samples 3,6,11
21	Unk9	831	0.355	Samples 1-3,5,6; Samples 4,7,9; Samples 8,10-12
22	Glutamic acid	988.5	0.349	Samples 1,3,6,7,9,11; Samples 2,8,10; Samples 4,5,12
23	Unk10	1175.25	0.324	Sample 1; Samples 2,5,6; Samples 3,4,7-12
24	Glycerol	663	0.303	Samples 1,4,12; Samples 2,3,5-7; Samples 8-11
25	Unk11	648	0.268	Samples 1-6; Samples 8,10,12; Samples 7,9,11
26	Threonine	777	0.247	Samples 1,5-7,9; Samples 2-4,8,10,11; Sample 12
27	5-Oxoproline	907.5	0.237	Samples 1,3,5,6,9,10; Samples 2,4,8,12; Samples 7,11
28	Unk12	1082.25	0.153	Samples 1,7,9,10; Samples 2,4,5,8,12; Samples 3,6,11
29	Unk13	984	0.136	Samples 1,3,9; Samples 2,4,8,10,12; Samples 5-7,11
30	Asparginine	1036.5	0.135	Samples 1,3,5-7,9; Samples 2,4,8,10,12; Sample 11
31	Unk14	781.5	0.120	Samples 1,6,7,9,11; Samples 2,5,8,12; Samples 3,4,10
32	Unk15	1030.5	0.112	Samples 1,3,9; Samples 2,4,5,10,12; Samples 6-8,11
33	Isoleucine	682.5	0.111	Samples 1,3,9; Samples 2,4,5,8,10,12; Samples 6,7,11
34	Unk16	1974.75	0.106	Samples 1-6; Samples 7,9,11; Samples 8,10,12
35	Unk17	902.25	0.100	Samples 1-6; Samples 7,9,11; Samples 8,10,12
36	Unk18	1360.5	0.0918	Samples 1,3,6; Samples 2,4,5; Samples 7-12
37	Methionine	897	0.0792	Samples 1-3,5,6; Samples 4,8,10,12; Samples 7,9,11
38	Unk19	798.75	0.0722	Samples 1,3,5-7,9; Samples 2,4,8,10,12; Sample 11
39	Leucine	654	0.0708	Samples 1,3,6; Samples 2,4,5; Samples 7-12
40	Phenylalanine	999	0.0642	Samples 1,3,6; Samples 2,4,5; Samples 7-12
41	Unk20	1188.75	0.064	Samples 1,3,6; Samples 2,4,5; Samples 7-12
42	Unk21	1140	0.0575	Samples 1,2,4-6; Samples 3,7-11; Sample 12
43	Unk22	1135.5	0.0558	Samples 1,3,6; Samples 2,4,5; Samples 7-12
44	Unk23	820.5	0.0548	Samples 1-6; Samples 7,9,11; Samples 8,10,12
45	Glutamine	1119	0.0542	Samples 1,3,6; Samples 2,4,5; Samples 7-12
46	Unk24	810	0.0520	Samples 1,3,6,7,9,11; Samples 2,4,8,10; Samples 5,12
47	Unk25	633	0.0488	Samples 1,6,7,9,12; Samples 2,4,5,8,10; Samples 3,11

48	*o*-Toluic acid	750	0.0468	Samples 1,6,7,9,11; Samples 2,4,8,12; Samples 3,5,10
49	Unk26	846	0.0448	Samples 1,3,6,7,9,11; Samples 2,4,8; Samples 5,10,12
50	Unk27	673.5	0.0435	Samples 1,3,6; Samples 2,4,5; Samples 7-12
51	Unk28	1136.25	0.0351	Samples 1-6; Samples 7,9,11; Samples 8,10,12
52	Unk29	1311	0.0312	Samples 1-6; Samples 7-10,12; Sample 11
53	Stearic acid	1443	0.0299	Samples 1-3,6,7,9,11; Samples 4,5,10,12; Sample 8

[a] The 3 out of 53 peaks shaded in teal have the same sample index assignment. The probability of this occurring by chance is 1.36×10^{-7}.
[b] The 3 out of 53 peaks shaded in blue have the same sample index assignment. The probability of this occurring by chance is 1.36×10^{-7}.
[c] The 2 out of 53 peaks shaded in yellow have the same sample index assignment. The probability of this occurring by chance is 1.79×10^{-6}.
[d] The 2 out of 53 peaks shaded in orange have the same sample index assignment. The probability of this occurring by chance is 4.44×10^{-5}.
[e] The 4 out of 53 peaks shaded in gray have the same sample index assignment. The probability of this occurring by chance is 2.50×10^{-12}.
[f] The 7 out of 53 peaks shaded in pink have the same sample index assignment. The probability of this occurring by chance is 2.09×10^{-22}.

Table C.8. List of *k*-means clustering results for the human metabolome data set using $k = 2$. Analytes shaded in green have matching sample index assignments.

Hit Number	Analyte Name	*RSD*[2]	Sample Index Assignments
1	1-Propanol	10.62	Samples 1-12,14-59; Sample 13
2	Acetone	10.22	Samples 1-16,18-59; Sample 17
3	*p*-Cresol	7.37	Samples 1-14,16-18,20-29,31-59; Samples 15,19,30
4	*o*-Cymene	6.63	Samples 1-4,6-31,33-42,44-57,59; Samples 5,32,43,58
5	Pentanoic acid, 4-methyl-	6.39	Samples 1-6,8,10,12,13,16-18,20-59; Samples 7,9,11,14,15,19
6	*p*-Cymene	6.16	Samples 1-4,6-57,59; Samples 5,58
7	2-Decanone	5.55	Samples 1-29,31-59; Samples 30
8	Propanoic acid, 2-methyl-	4.40	Samples 1,2,4-6,8,13,16-18,20-59; Samples 3,7,9-12,14,15,19
9	Phenol	4.33	Samples 1,3,7,9,11,13-15,19,23,26,27,29-31,44,58; Samples 2,4-6,8,10,12,16-18,20-22,24,25,28,32-43,45-57,59
10	*o*-Xylene	3.95	Samples 1-21,23,26,28-31,33-59; Samples 22,24,25,27,32
11	Phenol, 2,5-bis(1,1-dimethylethyl)-	3.87	Samples 1-32; Samples 33-59
12	Acetic acid ethenyl ester	3.23	Samples 1-12,14-17,19,20,22-59; Samples 13,18,21
13	Propanoic acid, ethyl ester	3.02	Samples 1-32; Samples 33-59
14	Octane, 3,3-dimethyl-	2.89	Samples 1-22,24-26,29,32-59; Samples 23,27,28,30,31
15	*n*-Propyl acetate	2.86	Samples 1-11,13-17,19,20,22-59; Samples 12,18,21
16	2,4-Dimethyl-1-heptene	2.65	Samples 1,3,7,11-18,28,37; Samples 2,4-6,8-10,19-27,29-36,38-59
17	Ethyl Acetate	2.43	Samples 1,4,5,7-9,11,14-16,18-20,22-59; Samples 2,3,6,10,12,13,17,21
18	Acetic acid	2.36	Samples 1-32; Samples 33-59
19	2,3-Pentanedione	2.33	Samples 1-4,6-15,17,19,20,22-46,48-57,59; Samples 5,16,18,21,47,58

20	Benzaldehyde	2.30	Samples 1-3,5-18,20,22,23,26,28,32-40,42,43,45-55,58,59; Samples 4,19,21,24,25,27,29-31,41,44,56,57
21	Ethylbenzene	2.24	Samples 1-24,27-31,33-57,59; Samples 25,26,32,58
22	1-Butanol	2.06	Samples 1,2,4,5,7-9,11,14,15,19,22,23,25-50,52-59; Samples 3,6,10,12,13,16-18,20,21,24,51
23	Benzaldehyde, 3-methyl-	1.76	Samples 1,2,4,7-9,15,19,20,22-51,53-57; Samples 3,5,6,10-14,16-18,21,52,58,59
24	Benzene, 4-ethenyl-1,2-dimethyl-	1.74	Samples 1-4,6-31,33-57,59; Samples 5,32,58
25	Propanoic acid	1.74	Samples 1-28; Samples 29-59
26	Ethanol	1.61	Samples 1-5,7-9,11,12,15,16,19-36,38,39,43-46,48-56,58,59; Samples 6,10,13,14,17,18,37,40-42,47,57
27	Styrene	1.60	Samples 1-18,20-22,24,29,32-59; Samples 19,23,25-28,30,31
28	Furfural	1.53	Samples 1-10,12-23,25-39,41-51,53-55,57-59; Samples 11,24,40,52,56
29	3-Furaldehyde	1.53	Samples 1,3,5,13,14,16-20,22,23,26-30,32-38,41-43,46,57; Samples 2,4,6-12,15,21,24,25,31,39,40,44,45,47-56,58,59
30	2-Propanol, 1-chloro-	1.48	Samples 1,2,4,5,7,9,11,12,14-16,19,20,22-59; Samples 3,6,8,10,13,17,18,21
31	1,2-Decanediol	1.45	Samples 1-32; Samples 33-59
32	Decane, 4-methyl-	1.32	Samples 1-22,24-26,29,32-35,37,39-41,43-46,48-59; Samples 23,27,38,30,31,36,38,42,47
33	*p*-Xylene	1.28	Samples 1,4,7,15,19,22,24,25,27,33,35,36,38-49,51,52,54-56,57,59; Samples 2,3,5,6,8-14,16-18,20,21,23,26,28-32,34,37,50,53,58
34	Pentane, 2,3,3-trimethyl-	1.22	Samples 1,3-6,10,12-14,16,17,20,21,30; Samples 2,7-9,11,15,18,19,22-29,31-59
35	Benzene, 1,4-dichloro-	1.21	Samples 1-32; Samples 33-59
36	Hexane, 3-methyl-	0.83	Samples 1,3,6,10,13,14,20,21,30; Samples 2,4,5,7-9,11,12,15-19,22-29,31-59
37	3-Pentenoic acid, 4-methyl-	0.83	Samples 1-3,5,6,8,10,12,13,17,20-44,46,48-52,54,57-59; Samples 4,7,9,11,14-16,18,19,45,47,53,55,56
38	1-Dodecanol, 3,7,11-trimethyl-	0.82	Samples 1,2,4-7,11,13-15,19,20,22,23,25-27,30,32-48,52,54-59; Samples 3,8-10,12,16-18,21,24,28,29,31,49-51,53
39	2-Butanol, 1-chloro-	0.78	Samples 1-3,5-8,10-15,17,18,20-22,24,28,37,43,47,51,53; Samples 4,9,16,19,23,25-27,29-36,38-42,44-46,48-50,52,54-59
40	Propanoic acid, propyl ester	0.78	Samples 1,4,6,10,14,20,21,23,25-27,30,31,37-39,41-45,50,51,54,55,57,59; Samples 2,3,5,7-9,11-13,15-19,22,24,28,29,32-36,40,46-49,52,53,56,58
41	Butane, 1,2,3,4-tetrachloro-	0.77	Samples 1-9,14,15,20,22-27,29,31,32,38,45,52-54,58; Samples 10-13,16-19,21,28,30,33-37,39,40-44,46-51,55-57,59
42	3-Decen-2-ol, (*E*)-	0.72	Samples 1,4-7,9,10,14,18,20-27,29-33,35,37,38,44,45,49,50,52-56,58; Samples 2,3,8,11-13,15-17,19,28,34,36,39-43,46-48,51,57,59
43	Benzene, 1,3-bis(1,1-dimethylethyl)-	0.60	Samples 1,3,4,7-18,21,24,25,28,29,34,35,37,39,43,45,49,51,53,58; Samples 2,5,6,19,20,22,23,26,27,30-33,36,38,40-42,44,46-48,50,52,54-57,59
44	Propane, 1,2,3-trichloro-2-methyl-	0.48	Samples 1,2,4-7,9-12,15-17,19,21,22,24,28,29,34,37-45,47-51,53-59; Samples 3,8,13,14,18,20,23,25-27,30-33,35,36,46,52
45	Propane, 1,1,3,3-tetrachloro-2-methyl-	0.46	Samples 1-3,8,9,13,14,18,33-36,38,39,41,43,46,47,49,52,55,57-59; Samples 4-7,10-12,15-17,19-32,37,40,42,44,45,48,50,51,53,54,56
46	Toluene	0.40	Samples 1,5,6,10-14,16-18,21,23,27,31,33,34,39,42,43,48,54,56,58,59; Samples 2-4,7-9,15,19,20,22,24-26,28-30,32,35-38,40,41,44-47,49-53,55,57
47	2-Propanol, 1,3-dichloro-	0.36	Samples 1-9,11-21,23,25-27,29-33,35,38-42,44,45,50,52,54-59; Samples 10,22,24,28,34,36,37,43,46-49,51,53
48	2-Propenoic acid	0.29	Samples 1-11,13-24,28,30,33,35,38,40-42; Samples 12,25-27,29,31,32,34,36,37,39,43-59

Table C.9. List of *k*-means clustering results for the human metabolome data set using $k = 3$. It was determined that none of the analytes had matching sample index assignments.

Hit Number	Analyte Name	RSD^2	Sample Index Assignments
1	1-Propanol	10.62	Samples 1,2,4-7,9-12,14-59; Samples 3,8; Sample 13
2	Acetone	10.22	Samples 1,2,4,5,7-9,11-15,19-24,26-40,42-59; Samples 3,6,10,16,18,25,41; Sample 17
3	*p*-Cresol	7.37	Samples 1,9,11,14,23,26,29,31; Samples 2-8,10,12,13,16-18,20-22,24,25,27,28,32-59; Samples 15,19,30
4	*o*-Cymene	6.63	Samples 1-4,6-8,10-19,21,23,30-34,36-38,40,42,46,47,49,51,52,54,56; Samples 5,58; Samples 9,20,22,24-29,35,39,41,43-45,48,50,53,55,57,59
5	Pentanoic acid, 4-methyl-	6.39	Samples 1-6,8,10,12,13,16-18,20-59; Samples 7,9,11,14,19; Sample 15
6	*p*-Cymene	6.16	Samples 1,3,4,6-8,10-16,18,19,21,31-33,36-38,40,42,47,49,51,52,54; Samples 2,9,17,20,22,23-30,34,35,39,41,43-46,48,50,53,55-57,59; Samples 5,58
7	2-Decanone	5.55	Samples 1-3,7-9,11-15,21,22,25,28,32-34,36-59; Samples 4-6,10,16-20,23,24,26,27,29,31,35,56; Sample 30
8	Propanoic acid, 2-methyl-	4.40	Samples 1,2,4-6,8,16-18,20-59; Samples 3,7,9-14,19; Sample 15
9	Phenol	4.33	Samples 1,3,7,9,11,13,14,19,23,26,27,29,30,31,44,58; Samples 2,4-6,8,10,12,16-18,20-22,24,25,28,32-43,45-57,59; Sample 15
10	*o*-Xylene	3.95	Samples 1-18,20,21,31,34,40,43,46-49,51,52,59; Samples 19,23,26,28-30,33,35-39,41,42,44,45,50,53-58; Samples 22,24,25,27,32
11	Phenol, 2,5-bis(1,1-dimethylethyl)-	3.87	Samples 1-32; Samples 33-37,39,46,47,51,55; Samples 38,40-45,48-50,52-54,56-59
12	Acetic acid ethenyl ester	3.23	Samples 1,3-11,13-17,19,20,24,25,34,35,45,48,49,50,53,54,56; Samples 2,22,23,26-33,36-44,46,47,51,52,55,57-59; Samples 12,18,21
13	Propanoic acid, ethyl ester	3.02	Samples 1,6-8,11,12,19; Samples 2-5,9,10,13,14-18,20-32; Samples 33-59
14	Octane, 3,3-dimethyl-	2.89	Samples 1-22,24-26,29,32-59; Samples 23,27,30,31; Sample 28
15	*n*-Propyl acetate	2.86	Samples 1,2,4,5,7,11,13,14,20,22,23,25-59; Samples 3,6,8-10,15-17,19,24; Samples 12,18,21
16	2,4-Dimethyl-1-heptene	2.65	Samples 1,3,7,11,14-18,20,28,37,38,42; Samples 2,4-6,8-10,19,21-27,29-36,39-41,43-59; Samples 12,13
17	Ethyl Acetate	2.43	Samples 1,5,8,9,20,23-36,38-41,43-59; Samples 2,3,6,10,12,13,21; Samples 4,7,11,14-19,22,37,42
18	Acetic acid	2.36	Samples 1,4,7,10,11,16,17,21-32; Samples 2,3,5,6,8,9,12-15,18-20; Samples 33-59
19	2,3-Pentanedione	2.33	Samples 1,3,4,10,20,22,23,26-28,30,31,33-35,37-39,43,50-54,56,57,59; Samples 2,6-9,11-15,17,19,24,25,29,32,36,40-42,44-46,48,49,55; Samples 5,16,18,21,47,58
20	Benzaldehyde	2.30	Samples 1-3,5-18,20,22,23,26,28,32-40,42,43,45-55,58,59; Samples 4,19,21,25,27,29-31,41,44,56,57; Sample 24
21	Ethylbenzene	2.24	Samples 1,2,4,5,7,15,19-24,27,28,30,33,35,36,38-43,45-49,51,52,55,57,59; Samples 25,26,32,58
22	1-Butanol	2.06	Samples 1,2,4,7,11,14,15,19,23,26-33,38,44-50,52-56,58,59; Samples 3,6,10,12,16-18,20,21,24,51; Samples 5,8,9,13,22,25,34-37,39,40-43,57
23	Benzaldehyde, 3-methyl-	1.76	Samples 1,2,4,7-9,15,19,20,22-38,40-51,54-56; Samples 3,12,13,17; Samples 5,6,10,11,14,16,18,21,39,52,53,57-59
24	Benzene, 4-ethenyl-1,2-dimethyl-	1.74	Samples 1-3,6-9,11,14-24,27,30,36,42,43,48,52,54; Samples 4,10,12,13,25,26,28,29,31,33-35,37-41,44-47,49-51,53,55-57,59; Samples 5,32,58

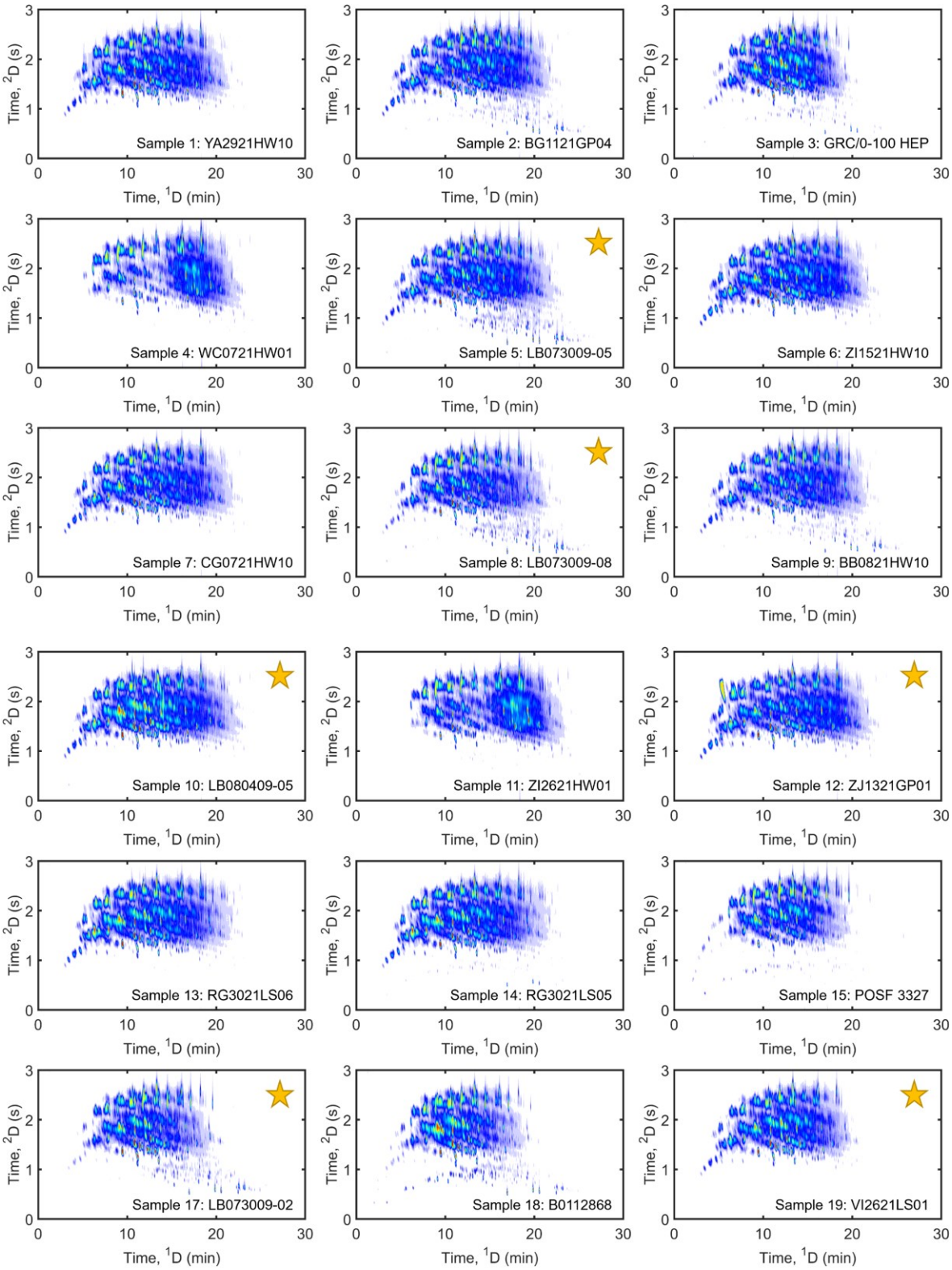
25	Propanoic acid	1.74	Samples 1,2,4,7,15,19; Samples 3,6,9-11,13,14,18,34,35,43,44,54,56,58; Samples 5,8,12,16,17,20-33,36-42,45-53,55,57,59
26	Ethanol	1.61	Samples 1,2,3,7,15,18-21,31,33-36,38,39,41,43-45,47,50-55,58,59; Samples 4,5,8,9,11,12,16,22-30,32,46,48,49,56; Samples 6,10,13,14,17,37,40,42,57
27	Styrene	1.60	Samples 1-4,6-8,10-18,20,21,33-43,45-49,51-55,59; Samples 5,9,22,24,29,32,44,50,56-58; Samples 19,23,25-28,30,31
28	Furfural	1.53	Samples 1,3,5,13,14,16-20,22,23,26-30,32-38,41-43,46,57; Samples 2,4,6-10,12,15,21,25,31,39,40,44,45,47-55,58,59; Samples 11,24,56
29	3-Furaldehyde	1.53	Samples 1,3,5,13,14,16-20,22,23,26-30,32-38,41-43,46,57-59; Samples 2,4,6-10,12,15,21,25,31,39,40,44,45,47-55; Samples 11,24,56
30	2-Propanol, 1-chloro-	1.48	Samples 1,4,7,9,15,19,22-28,30-36,38-42,44-50,52-59; Samples 2,5,11,12,14,16,20,29,37,43,51; Samples 3,6,8,10,13,17,18,21
31	1,2-Decanediol	1.45	Samples 1-5,8-13,15,16,19-32; Samples 6,7,14,17,18; Samples 33-59
32	Decane, 4-methyl-	1.32	Samples 1,4,5,8,9,15,19,20,22,29,33-36,39-41,43,45-49,51,52,54,55,57,59; Samples 2,3,6,7,10-14,16-18,21,24-26,32,37,44,50,53,56,58; Samples 23,27,28,30,31,38,42
33	*p*-Xylene	1.28	Samples 1,4,7,15,19,25,27,32-36,38-49,51,52,54-57,59; Samples 2,3,5,6,8-14,16-18,20,21,23,26,28-31,37,50,53,58; Samples 22,24
34	Pentane, 2,3,3-trimethyl-	1.22	Samples 1,4,5,10,12,14,17,21,30; Samples 2,7-9,11,15,18,19,22-29,31-59; Samples 3,6,13,16,20
35	Benzene, 1,4-dichloro-	1.21	Samples 1-32; Samples 33,37,39,40,42,43,46-55,58,59; Samples 34-36,38,41,44,45,56,57
36	Hexane, 3-methyl-	0.83	Samples 1,4,5,8-11,15-18,23,25,28,29,31,33,35,36,38,40-42,46,49,52,53; Samples 2,7,12,19,22,24,26,27,32,34,37,39,43-45,47,48,50,51,54-59; Samples 3,6,13,14,20,21,30
37	3-Pentenoic acid, 4-methyl-	0.83	Samples 1,2,5,6,8,10,12,13,17,20,29,33,35,37,42-44,46,48-52,54,58; Samples 3,21-28,30-32,34,36,38-41,57,59; Samples 4,7,9,11,14-16,18,19,45,47,53,55,56
38	1-Dodecanol, 3,7,11-trimethyl-	0.82	Samples 1,2,4-7,11,13-15,19,20,22,25-27,30,32-37,39-48,52,54-56,59; Samples 3,8-10,16-18,49,51,53; Samples 12,21,23,24,28,29,31,38,50,57,58
39	2-Butanol, 1-chloro-	0.78	Samples 1,2,5,7-9,11,12,14-16,20,22,24,28,36,43,47,48,51,53; Samples 3,6,10,13,17,18,21,37; Samples 4,19,23,25-27,29,30-35,38-42,44-46,49,50,52,54-59
40	Propanoic acid, propyl ester	0.78	Samples 1,4,6,10,14,20,23,25,27,30,31,37-39,41-43,45,50,51,54,55,57; Samples 2,3,5,8,9,11,15,16,18,19,24,28,32-36,40,46-49,52,53,56,58; Samples 7,12,13,17,21,22,26,29,44,59
41	Butane, 1,2,3,4-tetrachloro-	0.77	Samples 1-9,14,15,20,22-27,29,31,32,38,45,52-54,58; Samples 10,13,16-19,30; Samples 11,12,21,28,33-37,39-44,46-51,55-57,59
42	3-Decen-2-ol, (*E*)-	0.72	Samples 1,5,7,9,10,14,18,20,22-24,26,27,29-32,35,37,38,44,45,49,52-54,56,58; Samples 2,4,6,8,11,15,19,21,25,28,33,34,36,41,43,46-48,50,55,59; Samples 3,12,13,16,17,39,40,42,51,57
43	Benzene, 1,3-bis(1,1-dimethylethyl)-	0.60	Samples 1,3,7-9,11-18,21,24,25,28,29,58; Samples 2,4-6,10,19,20,22,32-39,41,43-46,48-56,59; Samples 23,26,27,30,31,40,42,47,57
44	Propane, 1,2,3-trichloro-2-methyl-	0.48	Samples 1,2,5,6,9,10,16,17,21,22,29,34,38,39,41-43,45,47-49,52,55,57-59; Samples 3,8,13,14,18,20,23,25-27,30-33,35,36,46; Samples 4,7,11,12,15,19,24,28,37,40,44,50,51,53-56
45	Propane, 1,1,3,3-tetrachloro-2-methyl-	0.46	Samples 1,8,9,14,33-36,39,41,46,47,49,52,55,59; Samples 2-7,13,15-18,20,38,40,42-45,48,51,54,56-58; Samples 10-12,19,21-32,37,50,53
46	Toluene	0.40	Samples 1,5,6,10-14,16-18,21,23,27,31,33,34,39,42,43,48,54,56,58,59; Samples 2,4,8,19,24,25,28,45; Samples 3,7,9,15,20,22,26,29,30,32,35-38,40,41,44,46,47,49-53,55,57
47	2-Propanol, 1,3-dichloro-	0.36	Samples 1-5,11-16,18-20,23,27,29-33,38,52,58; Samples 6-9,17,21,25,26,34-36,39-42,44-46,49,50,54-57,59; Samples 10,22,24,28,37,43,47,48,51,53

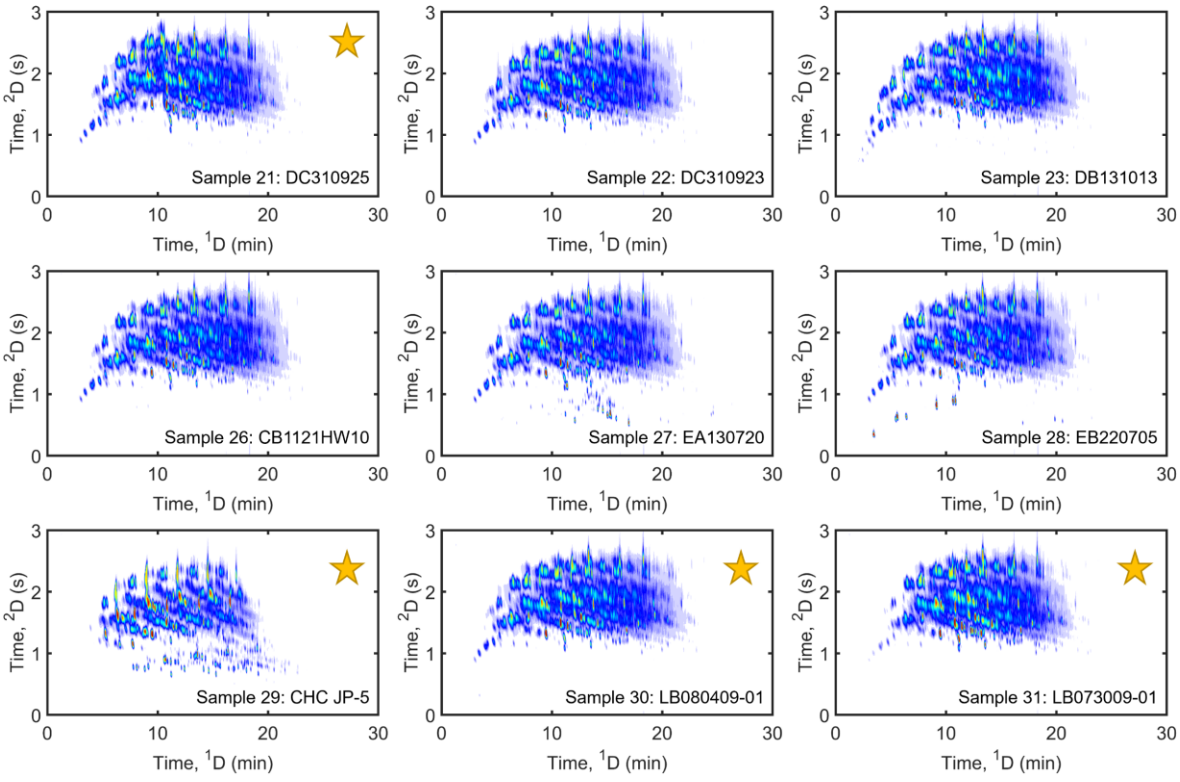
48	2-Propenoic acid	0.29	Samples 1-3,5-11,13,15,16,18,19,20,22,24,28,30,38,42; Samples 4,12,14,17,21,23,26,27,29,33-36,39-41,48-52,54,57-59; Samples 25,31,32,37,43-47,53,55,56

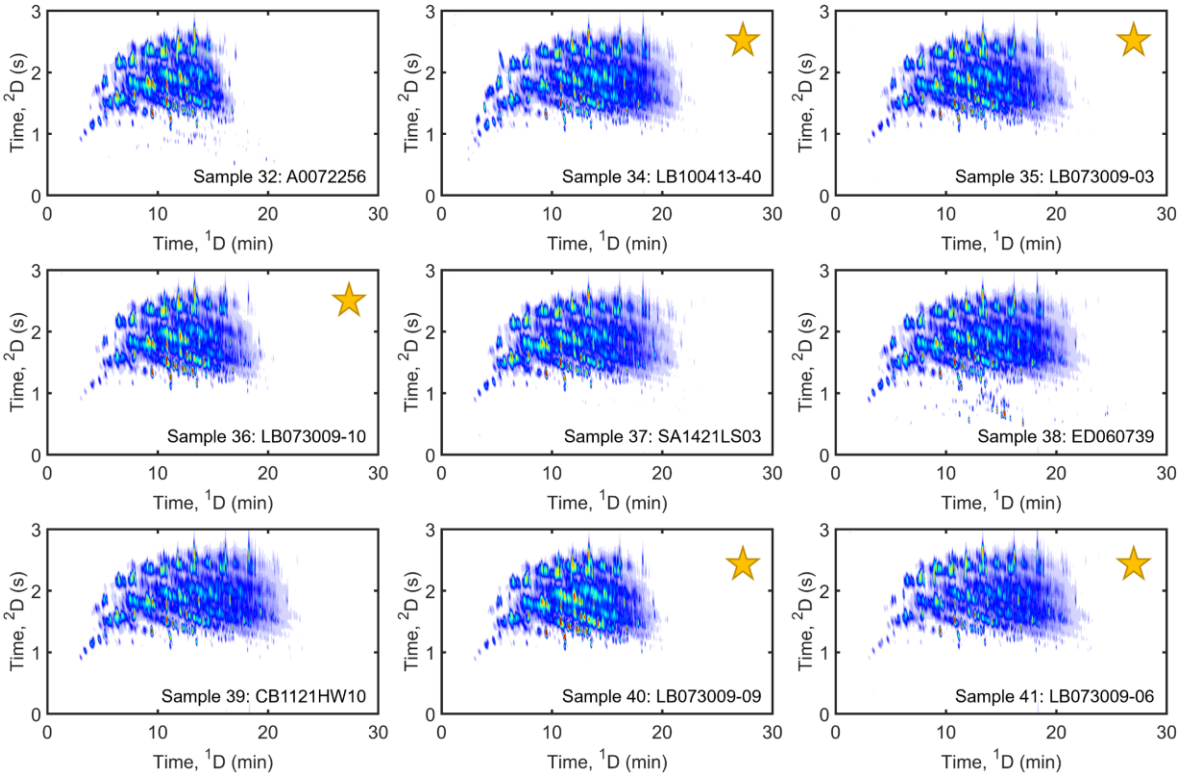
Appendix D

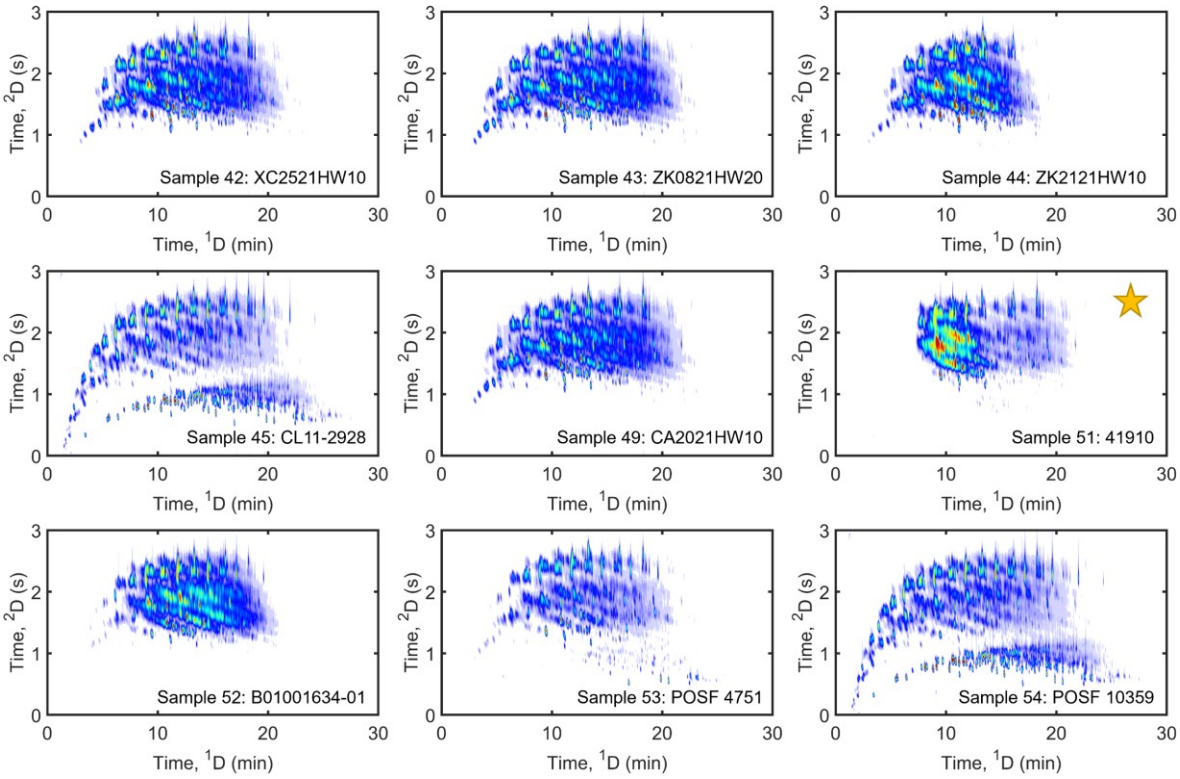
This appendix is reproduced from the Electronic Supplementary Material of C. N. Cain, G. S. Ochoa, R. E. Synovec, Enhancing Partial Least Squares Modeling of Comprehensive Two-Dimensional Gas Chromatography Time-of-Flight Mass Spectrometry Data by Tile-Based Variance Ranking, J. Chromatogr. A 1694 (2023), 463920.

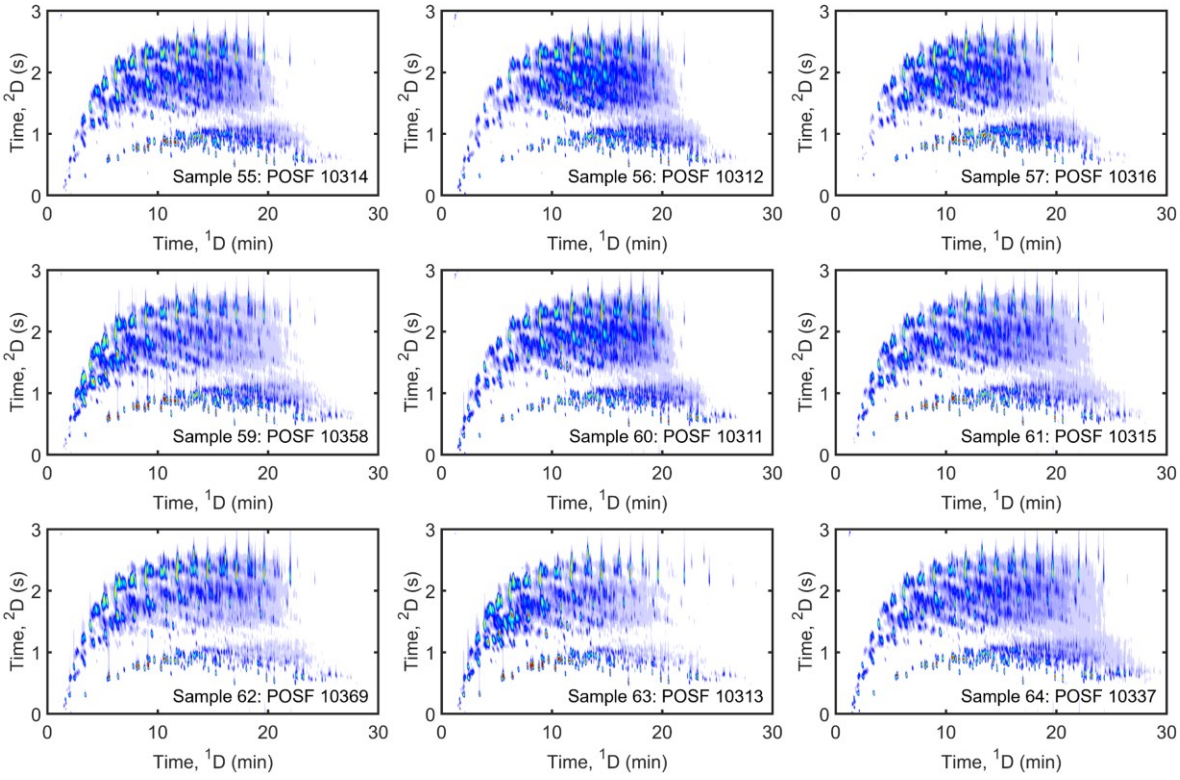
Note, the supplementary material begins on the next page.


Sample 1: YA2921HW10
Sample 2: BG1121GP04
Sample 3: GRC/0-100 HEP
Sample 4: WC0721HW01
Sample 5: LB073009-05
Sample 6: ZI1521HW10
Sample 7: CG0721HW10
Sample 8: LB073009-08
Sample 9: BB0821HW10
Sample 10: LB080409-05
Sample 11: ZI2621HW01
Sample 12: ZJ1321GP01
Sample 13: RG3021LS06
Sample 14: RG3021LS05
Sample 15: POSF 3327
Sample 17: LB073009-02
Sample 18: B0112868
Sample 19: VI2621LS01
Time, 1D (min)
Time, 2D (s)


Sample 21: DC310925
Sample 22: DC310923
Sample 23: DB131013
Sample 26: CB1121HW10
Sample 27: EA130720
Sample 28: EB220705
Sample 29: CHC JP-5
Sample 30: LB080409-01
Sample 31: LB073009-01
Time, ^{2}D (s)
Time, ^{1}D (min)


Sample 32: A0072256
Sample 34: LB100413-40
Sample 35: LB073009-03
Sample 36: LB073009-10
Sample 37: SA1421LS03
Sample 38: ED060739
Sample 39: CB1121HW10
Sample 40: LB073009-09
Sample 41: LB073009-06
Time, ^{2}D (s)
Time, ^{1}D (min)


Sample 42: XC2521HW10
Sample 43: ZK0821HW20
Sample 44: ZK2121HW10
Sample 45: CL11-2928
Sample 49: CA2021HW10
Sample 51: 41910
Sample 52: B01001634-01
Sample 53: POSF 4751
Sample 54: POSF 10359
Time, 2D (s)
Time, 1D (min)


Sample 55: POSF 10314
Sample 56: POSF 10312
Sample 57: POSF 10316
Sample 59: POSF 10358
Sample 60: POSF 10311
Sample 61: POSF 10315
Sample 62: POSF 10369
Sample 63: POSF 10313
Sample 64: POSF 10337
Time, 2D (s)
Time, 1D (min)

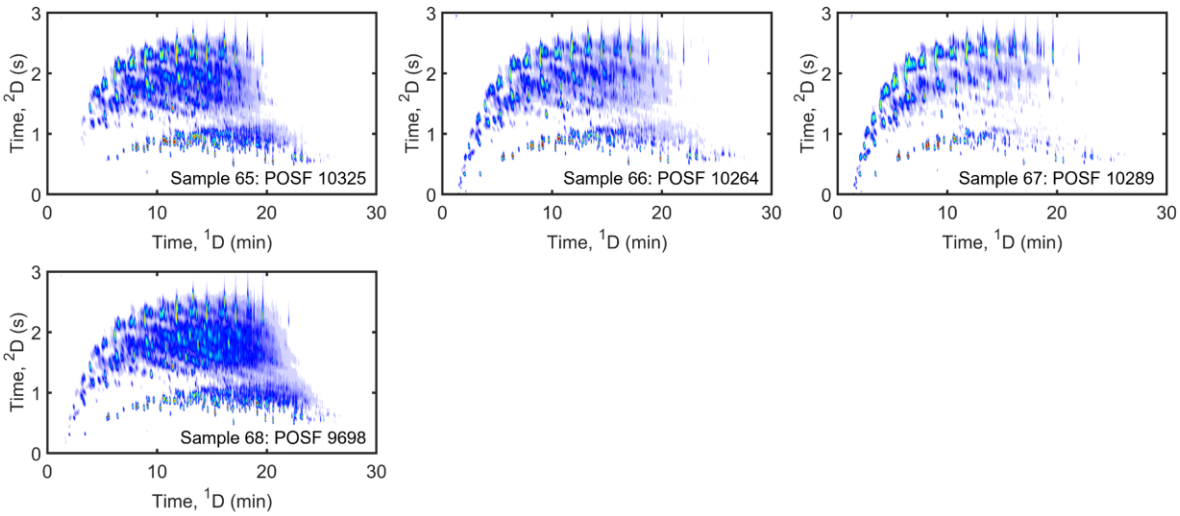


Figure D.1. Total ion current (TIC) GC×GC-TOFMS chromatograms of the 58 fuels used in this study. The 16 fuels used for the external validation set are marked by a yellow star. The original sample numbers and names from Berrier et al. [1] were kept for consistency.

[1] K.L. Berrier, C.E. Freye, M.C. Billingsley, R.E. Synovec, Predictive Modeling of Aerospace Fuel Properties Using Comprehensive Two-Dimensional Gas Chromatography with Time-Of-Flight Mass Spectrometry and Partial Least Squares Analysis, Energy Fuels 34 (2020) 4084–4094. https://doi.org/10.1021/acs.energyfuels.9b04108.

Table D.1. Analytes highly loaded in the PLS model for viscosity, which was built using all the features discovered by tile-based variance ranking (Figure 6.4D).

LRV Rank	Time, ^{1}D (min)	Time, ^{2}D (s)	*RSD*^{2}	Top *m/z*	Analyte	MV
Positive						
1	1.15	2.76	12.03	46	Ethanol	864
2	9.1	0.39	11.31	88	Ethanol, 2-(2-methoxyethoxy)-	967
3	15.8	2.16	0.12	166	1-Eicosanol	813
4	9.45	0.32	1.76	45	Unknown	
5	10.15	2.32	0.07	170	Decane, 4-ethyl-	914
6	1.25	2.97	4.69	58	n-Hexane	917
7	16.2	2.62	0.16	212	Dodecane, 2,6,10-trimethyl-	949
8	26.5	0.55	7.81	57	Unknown	
9	16.1	2.31	0.17	155	Tridecane, 2-methyl-	889
10	23.8	2.52	1.6	114	Pentadecane, 2,6,10,14-tetramethyl-	899
Negative						
1	2.8	0.45	2.46	81	Cyclohexene, 3-methyl-	829
2	20	0.72	4.09	168	Unknown	
3	6	0.62	7.35	101	Thiophene, tetrahydro-2,5-dimethyl-, trans-	832
4	6.15	0.62	5.9	101	Thiophene, tetrahydro-2,5-dimethyl-, trans-	832
5	19.3	0.74	3.15	59	Biphenylene, 1,2,3,6,7,8,8a,8b-octahydro-, trans-	801
6	6.85	1.82	0.22	140	Cyclohexane, 1,5-diethyl-2,3-dimethyl-	835
7	13.1	1.61	0.64	149	4-Tridecyne	810
8	14.1	1.58	1.14	133	3,5-Octadiene, 4,5-diethyl-	803
9	22	2.16	2.58	226	Hexadecane	932
10	22	2.23	2.67	226	Hexadecane	945

Table D.2. Analytes highly loaded in the PLS model for hydrogen content, which was built using all the features discovered by tile-based variance ranking (Figure 6.4E).

LRV Rank	Time, ^{1}D (min)	Time, ^{2}D (s)	*RSD*^{2}	Top *m/z*	Analyte	MV
Positive						
1	25.75	1.47	7.99	83	Cyclohexane, 1,1"-(1-methyl-1,3-propanediyl)bis-	897
2	20	0.72	4.09	168	Unknown	
3	19.3	0.74	3.15	59	Biphenylene, 1,2,3,6,7,8,8a,8b-octahydro-, trans-	801
4	14.1	1.58	1.14	133	3,5-Octadiene, 4,5-diethyl-	803
5	6	0.62	7.35	101	Thiophene, tetrahydro-2,5-dimethyl-, trans-	832
6	6.15	0.62	5.9	101	Thiophene, tetrahydro-2,5-dimethyl-, trans-	832
7	14.7	2.45	0.13	114	Tetradecane	907
8	15.8	2.16	0.12	166	1-Eicosanol	813
9	13.15	2.47	0.42	167	Nonadecane	855
10	10.15	2.32	0.07	170	Decane, 4-ethyl-	914
Negative						
1	1.45	0.14	3.63	69	Hexane, 2-methyl-	905
2	1.7	0.06	3.56	84	Cyclohexane	855
3	20.8	0.76	4.03	165	3,5-Dimethyl-3-phenyl-3H-pyrazole	805
4	1.25	2.97	4.69	58	*n*-Hexane	917
5	24.1	0.62	2.75	125	Naphthalene, 2-(1-methylethyl)-	801
6	1.4	0.14	2.96	85	Hexane, 2-methyl-	925
7	6.65	1.11	0.32	78	Pentalene, octahydro-1-methyl-	906
8	18.4	0.66	1.83	144	Naphthalene, 1,2-dihydro-6-methyl-	806
9	1.35	0.01	2.99	85	Pentane, 2,2-dimethyl-	916
10	24.35	0.51	2.14	113	Cyclopentane, 1-butyl-2-propyl-	819

Table D.3. Analytes highly loaded in the PLS model for heat of combustion, which was built using all the features discovered by tile-based variance ranking (Figure 6.4F).

LRV Rank	Time, ^{1}D (min)	Time, ^{2}D (s)	*RSD*^{2}	Top *m/z*	Analyte	MV
Positive						
1	25.75	1.47	7.99	83	Cyclohexane, 1,1"-(1-methyl-1,3-propanediyl)bis-	897
2	10.15	2.32	0.07	170	Decane, 4-ethyl	914
3	11.65	2.44	0.47	170	Tetradecane	895
4	14.7	2.45	0.13	114	Tetradecane	907
5	11.8	2.31	0.49	171	Dodecane	947
6	11.8	2.47	0.6	171	Dodecane	951
7	14.1	1.58	1.14	133	3,5-Octadiene, 4,5-diethyl-	803
8	6.1	2.02	0.81	100	Decane	958
9	9	2.34	0.77	157	Undecane	954
10	8	2.27	0.07	86	Decane, 3-methyl-	961
Negative						
1	1.35	0.01	2.99	85	Pentane, 2,2-dimethyl-	916
2	13.1	1.61	0.64	149	4-Tridecyne	810
3	1.7	0.06	3.56	84	Cyclohexane	855
4	1.45	0.14	3.63	69	Hexane, 2-methyl-	905
5	1.25	2.97	4.69	58	*n*-Hexane	917
6	1.4	0.14	2.96	85	Hexane, 2-methyl-	925
7	9.1	0.39	11.31	88	Ethanol, 2-(2-methoxyethoxy)-	967
8	4.85	1.46	0.34	125	Cyclohexane, 1,1,2,3-tetramethyl-	849
9	1.65	0.21	2.73	100	Cyclopentane, 1,3-dimethyl-, cis-	880
10	20.8	0.76	4.03	165	3,5-Dimethyl-3-phenyl-3H-pyrazole	805

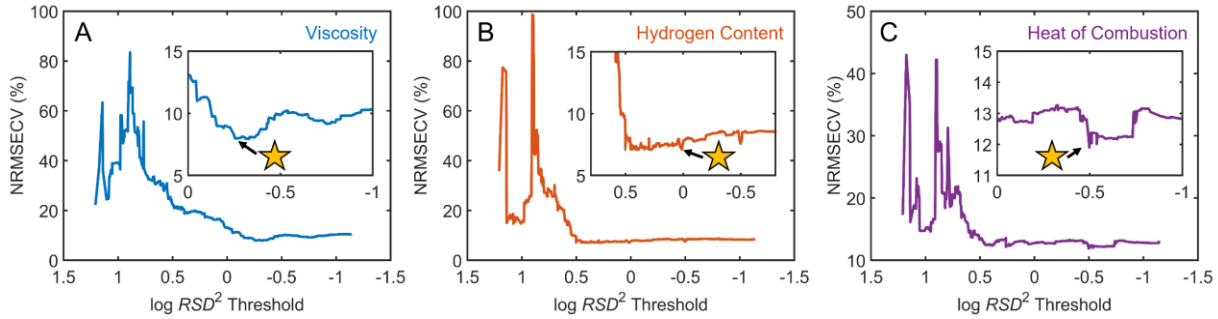


Figure D.2. Selection of an RSD^2 threshold to improve the performance of PLS for modeling (A) viscosity, (B) hydrogen content, and (C) heat of combustion. These plots show the NRMSECV for the PLS model as a function of the log RSD^2. The arrow and yellow star indicate the threshold that showed the smallest modeling error. (A) The PLS model built using the top 420 features discovered ($RSD^2 > -0.25$) has the lowest NRMSECV of 7.94 %. (B) The PLS model built using the top 378 features discovered ($RSD^2 > 0.11$) has the lowest NRMSECV of 7.10 %. (C) The PLS model built using the top 464 features discovered ($RSD^2 > -0.50$) has the lowest NRMSECV of 11.90 %.

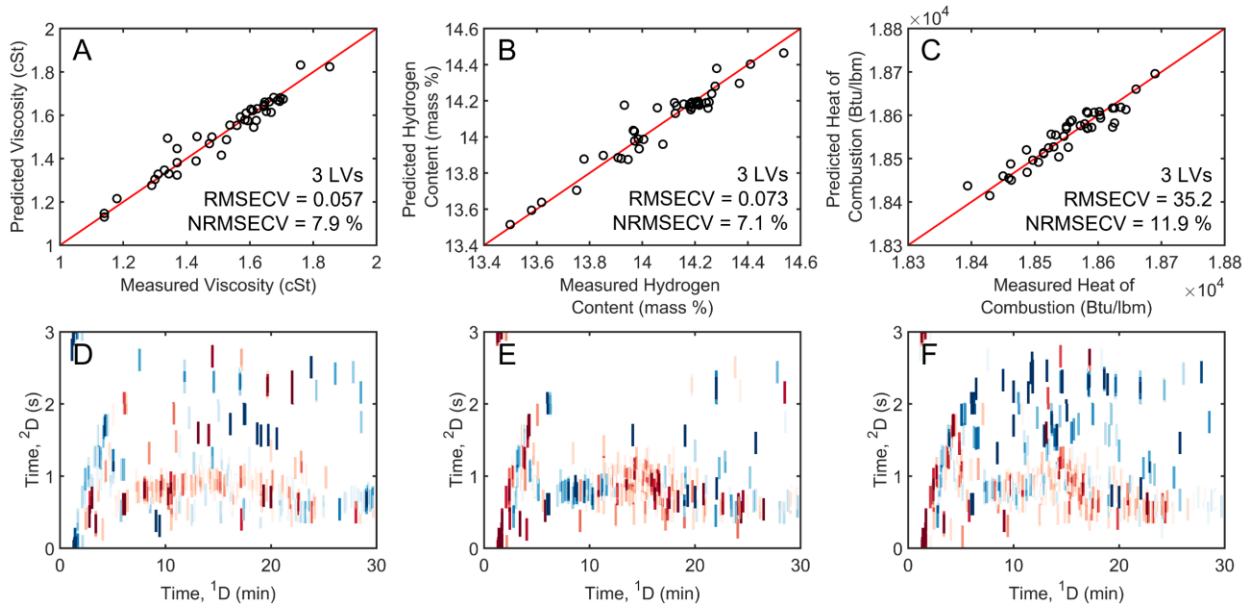


Figure D.3. PLS prediction of viscosity (left), hydrogen content (middle), and heat of combustion (right) after placing a threshold on the RSD^2, which is seen in Figure D.2. (A-C) The red line represents ideal agreement between the predicted and measured values. The number of LVs, RMSECV, and NRMSECV for each PLS model is provided. (D-F) LRVs generated for each physical property, where positively loaded values are highlighted in blue and negatively loaded values are shown in red. All LRVs are plotted on the same color scale.

Table D.4. Analytes highly loaded in the final PLS model for viscosity, which was built using the features selected by RReliefF feature optimization (Figure 6.7D).

LRV Rank	Time, ^{1}D (min)	Time, ^{2}D (s)	*RSD*^{2}	Top *m/z*	Analyte	MV
Positive						
1	1.25	2.97	4.69	58	Hexane	917
2	15.8	2.16	0.12	166	1-Eicosanol	813
3	23.8	2.52	1.6	114	Pentadecane, 2,6,10,14-tetramethyl-	899
4	9.45	0.32	1.76	45	Unknown	
5	5.25	1.59	0.63	82	Cyclooctane, methyl-	811
6	9.1	0.39	11.31	88	Ethanol, 2-(2-methoxyethoxy)-	967
7	18.7	1.6	0.89	203	(-)-Neoclovene-(I), dihydro-	801
8	28.55	1.76	2.32	57	Cyclohexane, hexyl-	802
9	5.8	0.53	3.53	106	*p*-Xylene	878
10	10.15	2.32	0.07	170	Decane, 4-ethyl-	914
Negative						
1	2.8	0.45	2.46	81	Cyclohexene, 1-methyl-	829
2	6	0.62	7.35	101	Thiophene, tetrahydro-2,5-dimethyl-, trans-	832
3	6.15	0.62	5.9	101	Decane, 2,3,5,8-tetramethyl-	833
4	6.9	2.32	0.08	98	*cis*-2,3-Dimethylthiophane	802
5	5.85	1.86	0.17	139	Unknown	
6	20	0.72	4.09	168	3-Undecene, (*Z*)-	836
7	8.25	0.82	2.78	60	Naphthalene, 2,6-dimethyl-	829
8	14.45	2.67	1.03	184	Tetradecane	927
9	22.5	0.62	2.05	120	Benzene, 1,2,3-trimethyl-	922
10	19.3	0.74	3.15	59	Unknown	

Table D.5. Analytes highly loaded in the final PLS model for hydrogen content, which was built using the features selected by RReliefF feature optimization (Figure 6.7E).

LRV Rank	Time, ^{1}D (min)	Time, ^{2}D (s)	*RSD*^{2}	Top *m/z*	Analyte	MV
Positive						
1	25.75	1.47	7.99	83	Cyclohexane, 1,1'-(1-methyl-1,3-propanediyl)bis-	892
2	19.3	0.74	3.15	59	Biphenylene, 1,2,3,6,7,8,8a,8b-octahydro-, trans-	801
3	20	0.72	4.09	168	Unknown	
4	14.1	1.58	1.14	133	3,5-Octadiene, 4,5-diethyl-	803
5	6	0.62	7.35	101	Thiophene, tetrahydro-2,5-dimethyl-, trans-	832
6	6.15	0.62	5.9	101	Thiophene, tetrahydro-2,5-dimethyl-, trans-	832
7	2.8	0.45	2.46	81	Cyclohexene, 1-methyl-	829
8	22.05	0.84	3.84	174	1H-Indene, 2,3-dihydro-1,1,5,6-tetramethyl-	840
9	5.15	0.82	2.44	110	Bicyclo[2.2.2]octane	886
10	19	0.6	2.85	149	1,3-Cyclohexadiene, 1,2,3,4,5,6-hexamethyl-	820
Negative						
1	16.2	0.7	2	129	Benzene, (2-cyclopropylethenyl)-	804
2	19	0.44	2.55	120	Unknown	
3	17.2	0.41	1.11	102	Naphthalene	834
4	9.1	0.39	11.31	88	Ethanol, 2-(2-methoxyethoxy)-	967
5	24.35	0.51	2.14	113	Naphthalene, 2,6-dimethyl-	926
6	1.7	0.06	3.56	84	Cyclohexane	890
7	1.45	0.14	3.63	69	Hexane, 2-methyl-	905
8	13.8	0.75	2.94	117	Benzene, 1-methyl-4-(2-propenyl)-	947
9	8.55	0.65	0.92	91	Benzene, 1,2,3-trimethyl-	871
10	3.4	0.37	2.82	94	Toluene	956

Table D.6. Analytes highly loaded in the final PLS model for heat of combustion, which was built using the features selected by RReliefF feature optimization (Figure 6.7F).

LRV Rank	Time, ^{1}D (min)	Time, ^{2}D (s)	*RSD*^{2}	Top *m/z*	Analyte	MV
Positive						
1	11.65	2.44	0.47	170	Dodecane	895
2	25.75	1.47	7.99	83	Cyclohexane, 1,1'-(1-methyl-1,3-propanediyl)bis-	892
3	14.1	1.58	1.14	133	3,5-Octadiene, 4,5-diethyl-	803
4	14.7	2.45	0.13	114	Tetradecane	907
5	10.15	2.32	0.07	170	Decane, 4-ethyl-	914
6	11.8	2.47	0.6	171	Dodecane	951
7	11.8	2.31	0.49	171	Dodecane	947
8	20	0.72	4.09	168	Unknown	
9	6.1	2.02	0.81	100	Decane	958
10	13.15	2.47	0.42	167	Tridecane	855
Negative						
1	9.1	0.39	11.31	88	Ethanol, 2-(2-methoxyethoxy)-	967
2	1.35	0.01	2.99	85	Pentane, 2,2-dimethyl-	916
3	1.4	0.14	2.96	85	Hexane, 2-methyl-	925
4	1.7	0.06	3.56	84	Cyclohexane	890
5	1.25	2.97	4.69	58	*n*-Hexane	917
6	1.7	0.37	2.73	57	Butane, 2,2,3,3-tetramethyl-	877
7	9.45	0.32	1.76	45	Unknown	
8	1.65	0.21	2.73	100	Cyclopentane, 1,3-dimethyl-, cis-	880
9	22.6	0.51	2.94	134	Benzo[b]thiophene, 7-ethyl-	816
10	1.45	0.14	3.63	69	Hexane, 2-methyl-	905

Appendix E

This appendix is reproduced from the Electronic Supplementary Material of C. N. Cain, T. J. Trinklein, G. S. Ochoa, R. E. Synovec, Tile-Based Pairwise Analysis of GC×GC-TOFMS Data to Facilitate Analyte Discovery and Mass Spectrum Purification, Anal. Chem. 94 (2022), 5658-5666.

Experimental Conditions

Spiked diesel fuel. Diesel fuel was spiked with a non-native internal standard, 1,2,3-trichlorobenzene (550 ppm), and 18 non-native analytes, which are listed in Table E.1. These 18 analytes were equally divided into three groups: analytes at ~10/20 ppm, analytes at ~30/60 ppm, and analytes at ~100/200 ppm (Table E.1). Therefore, the sample for class 1 was composed of the spiked analytes at 10, 30, or 100 ppm while the sample for class 2 was composed of the analytes at either 20, 60, or 200 ppm. A neat standards solution was also created by adding 200 µL of each compound in a scintillation vial. The standards solution was used to create a library to accurately measure the retention times of each spiked analyte and identify them by their mass spectrum. Six replicates of both spiked diesel classes and two replicates of the neat standards mixture (14 total chromatograms) were collected with the Pegasus BT 4D GC×GC-TOFMS instrument (LECO Corporation, St. Joseph, MI, USA) equipped with an L-PAL3 GC autosampler, Agilent 7890 GC (Agilent Technologies, Palo Alto, CA, USA), and a quad-jet thermal modulator. A 0.5 µL aliquot of the spiked diesel solutions was delivered to the inlet with a split ratio of 100:1. For the neat standards mixture, a 0.1 µL aliquot was injected into the inlet with a split ratio of 300:1. The inlet temperature was held constant at 275 ºC. The ^{1}D column was a polar Rxi-17Sil MS (30 m × 250 µm i.d. × 0.25 µm; Restek, Bellefonte, PA, USA) and the ^{2}D column was a non-polar Rxi-1 MS (2 m × 180 µm × 0.18 µm; Restek). The ^{1}D column was held at 40 ºC for 1.5 min before ramping to 250 ºC at 5 ºC/min, where it was held for 2 min. The same temperature program was used for the ^{2}D oven and modulator with an offset of +5 ºC and +25 ºC, respectively. The carrier gas, ultra-high purity helium (Grade 5, 99.999%, Praxair, Seattle, WA, USA), operated at a constant flow rate of 2 mL/min. The ^{1}D effluent was reinjected on the ^{2}D column at a P_M of 4 s with a hot pulse of 0.4 s and a cold pulse of 1.6 s. The ion source and transfer line temperatures were set to 225 ºC and 285 ºC, respectively. The TOFMS collected *m/z* 45-300 at 100 Hz with an electron ionization energy of 70 eV after a 10 s acquisition delay.

Cacao bean study. Cacao beans from the Ivory Coast were sourced by Theo Chocolate (Seattle, WA, USA), which all initially had 0 % mold coverage. A subset of the cacao beans was added to filtered water, where they molded to 100 % coverage. Hence, these samples made up the unmolded (0 % coverage) and molded (100 % coverage) classes for tile-based 1v1 analysis. The headspace of these coffee beans was extracted using a 65 µm PDMS/DVB solid-phase microextraction (SPME) fiber at 60 ºC for 10 min. Separations of the headspace extracts were performed on GC×GC-TOFMS configured with an Agilent 6890N GC (Agilent Technologies) and a Pegasus III TOFMS (LECO Corporation). The SPME fiber was desorbed in the inlet at 250 ºC for 5 min. The ^{1}D column was a RTX-5MS (20 m × 250 µm × 0.5 µm; Restek) while the ^{2}D column was a RTX-200MS (2 m × 180 µm × 0.2 µm; Restek). The He flow rate was constant at 1 mL/min. The ^{1}D column was held at 40 ºC for 5 min before ramping to 140 ºC at 8 ºC/min

and then ramping to 250 ºC at 30 ºC/min. The same temperature program was used for the ^{2}D oven and modulator with an offset of 10 ºC and 20 ºC, respectively. The P_M was 1.5 s. The ion source and transfer line temperatures were set to 250 ºC and 280 ºC, respectively. The TOFMS collected *m/z* 40-250 at 100 Hz with an electron ionization energy of 70 eV. This data set was previously collected, and additional information regarding the experimental parameters can be found in previous publications [1,2].

[1] E.M. Humston, Y. Zhang, G.F. Brabeck, A. McShea, R.E. Synovec, Development of a GCxGC-TOFMS method using SPME to determine volatile compounds in cacao beans, J. Sep. Sci. 32 (2009) 2289–2295. https://doi.org/10.1002/jssc.200900143.

[2] E.M. Humston, J.D. Knowles, A. McShea, R.E. Synovec, Quantitative assessment of moisture damage for cacao bean quality using two-dimensional gas chromatography combined with time-of-flight mass spectrometry and chemometrics, J. Chromatogr. A. 1217 (2010) 1963–1970. https://doi.org/10.1016/j.chroma.2010.01.069.

Table E.1. Actual concentrations and retention times for the 18 spiked analytes. Group number refers to if the analyte was designated as a 10/20 ppm spike (Group #1), 30/60 ppm spike (Group #2), or 100/200 ppm spike (Group #3). Within each group, the compounds are sorted in ascending order of their retention time on ^{1}D.

Group	Compound	Actual Concentration (ppm)		^1t_R (min)	^2t_R (s)
		Class 1	Class 2		
1	2-methylthiophene	9.8	19.4	3.7	0.14
1	α-pinene	11.3	22.4	5.4	1.44
1	1-bromooctane	10.9	21.6	12.3	1.45
1	1,6-dichlorohexane	10.3	20.5	14.1	0.72
1	phenylethyl alcohol	10.2	20.2	14.7	0.45
1	2-dodecanone	10.8	21.4	19.5	1.40
2	2-heptanone	34.4	68.3	6.1	0.59
2	butyl sulfide	33.0	65.6	10.7	1.42
2	methyl salicylate	34.2	68.0	16.5	0.62
2	methyl decanoate	36.1	71.7	17.5	1.40
2	citral	33.8	67.2	17.9	0.77
2	dibenzylamine	40.3	80.2	30.5	0.58
3	methyl caproate	123.6	245.8	6.7	0.76
3	1-bromoheptane	107.0	212.8	9.4	1.26
3	1-nonanol	118.4	235.4	13.4	1.32

3	2-undecanone	116.8	232.2	17.0	1.30
3	diphenyl sulfide	106.3	211.4	27.9	0.51
3	dibutyl phthalate	115.2	229.0	34.9	0.64

Class Comparison Enabled-Mass Spectrum Purification (CCE-MSP) Theory

Analyte identification of the discovered hits was first performed with CCE-MSP. The computational principles of CCE-MSP will be briefly discussed herein, and the reader is directed to our previous report for more details [3]. For a given hit, its spectrum is comprised of contributions from the analyte changing between classes and constant background interferences. The concepts can be expressed beginning with [3],

$$S(m/z)_1 = k_1\left[R(m/z)_{\mathrm{A}}\, C_{\mathrm{A},1} + \sum_{i=1}^{n} R(m/z)_{\mathrm{Int}} C_{\mathrm{Int},1}\right] \quad \text{(E.1)}$$

$$S(m/z)_2 = k_2\left[R(m/z)_{\mathrm{A}}\, C_{\mathrm{A},2} + \sum_{i=1}^{n} R(m/z)_{\mathrm{Int}} C_{\mathrm{Int},2}\right] \quad \text{(E.2)}$$

where $S(m/z)$ is the hit spectrum, $R(m/z)_{\mathrm{A}}$ is the analyte spectrum, C_{A} is the analyte concentration, $R(m/z)_{\mathrm{Int}}$ is the interference spectrum, and C_{Int} is the interferent concentration for class 1 and 2, respectively. Even if the total data set was normalized prior to non-targeted analysis, small differences in the background spectrum for each class generally exists (i.e., $k_1 \neq k_2$) [3]. Therefore, the hit spectra are normalized to a "pure" interferent *m/z* prior to subtracting Eq. E.1 from Eq. E.2 [3].

For tile-based 1v1 analysis, we propose that pure interferent *m/z* can be discovered using the *RM* and the previously established *LOF* metric [4]. A small *RM* can identify potential interferent *m/z* since it measures the relative difference in signals between classes. The *LOF* metric can then be used to determine *m/z* purity by describing the similarities/differences in the peak shapes between classes. Briefly, the *LOF* is defined as [4]

$$LOF(m/z) = \sqrt{\frac{\sum_i (y-x)_i^2}{\sum_i x_i^2}} \times 100 \quad \text{(E.3)}$$

where x and y are the summed ^{2}D peak profiles for each class. Hence, a *m/z* with a sufficiently low *RM* and *LOF* indicates a pure interferent *m/z* because the peak signals and shapes are consistent between classes.

Using a *m/z* identified to be due the background interferences, the normalization factor (k_1/k_2) is calculated as [3]

$$\frac{k_1}{k_2} = \frac{s(m/z)_{\mathrm{Int},1}}{s(m/z)_{\mathrm{Int},2}} \quad \text{(E.4)}$$

where $s(m/z)_{\mathrm{Int},1}$ and $s(m/z)_{\mathrm{Int},2}$ are the signals in each class. This normalization factor is applied to $S(m/z)_2$ in Eq. E.2 to reduce the contributions from the interference spectrum. CCE-MSP then obtains the pure analyte spectrum ($R(m/z)_{\mathrm{A}}$) as [3]

$$\left(\frac{k_1}{k_2}\right) S(m/z)_2 - S(m/z)_1 = k_1 R(m/z)_{\mathrm{A}} \left(C_{\mathrm{A},2} - C_{\mathrm{A},1}\right) \quad \text{(E.5)}$$

Note, while $R(m/z)_{\mathrm{A}}$ is scaled by $k_1(C_{\mathrm{A},2}\text{-}C_{\mathrm{A},1})$, this scaling does not affect analyte identification via mass spectrum matching to library spectra [3]. If multiple interferent *m/z* are identified with the *RM* and *LOF* signal consistency metrics for tile-based 1v1 analysis, then the calculations in

Eqs. E.4 and E.5 are repeated for each *m/z*. The resulting pure analyte spectra are then averaged together and used for identification.

While the interferent *m/z* are used for spectrum normalization in standard CCE-MSP [3], knowledge of these *m/z* can also improve MCR-ALS decomposition. Generally, MCR-ALS is performed using all *m/z* to resolve every peak, but inclusion of every noisy and/or interferent *m/z* hinders its ability to resolve the target analyte. CCE-MSP assisted MCR-ALS overcomes this problem by only performing MCR-ALS on the *m/z* not identified as a pure interferent *m/z*.

[3] G.S. Ochoa, P.E. Sudol, T.J. Trinklein, R.E. Synovec, Class comparison enabled mass spectrum purification for comprehensive two-dimensional gas chromatography with time-of-flight mass spectrometry, Talanta. 236 (2022) 122844. https://doi.org/10.1016/j.talanta.2021.122844.

[4] Y. Izadmanesh, E. Garreta-Lara, J.B. Ghasemi, S. Lacorte, V. Matamoros, R. Tauler, Chemometric analysis of comprehensive two dimensional gas chromatography–mass spectrometry metabolomics data, J. Chromatogr. A. 1488 (2017) 113–125. https://doi.org/10.1016/j.chroma.2017.01.052.

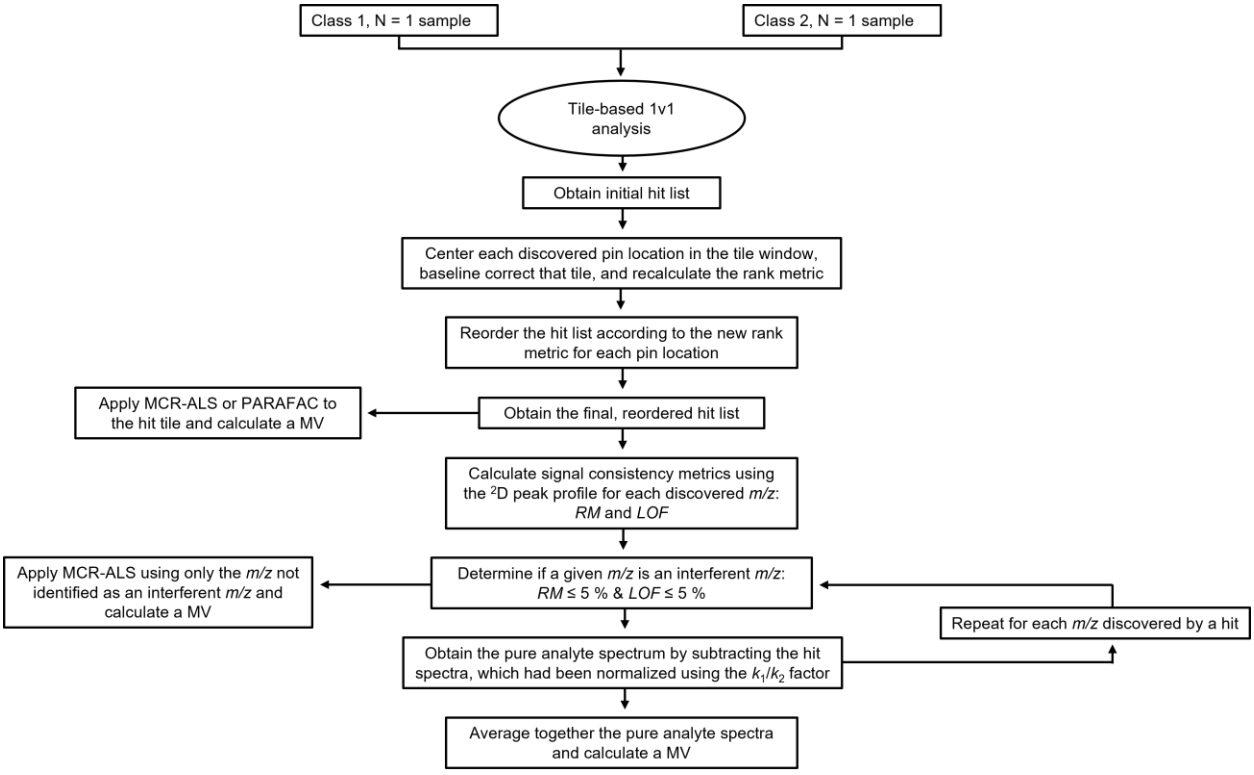


Figure E.1. Flow chart depicting the workflow designed in this study to discover analytes via tile-based pairwise (1v1) analysis and identify them using either standard chemometric methods (MCR-ALS and PARAFAC), CCE-MSP, or CCE-MSP assisted MCR-ALS.

Table E.2. Tile-based 1v1 analysis and F-ratio analysis hit lists for the spiked diesel comparison. *RM* for 1v1 analysis were generated using paired replicate chromatograms from class 1 and class 2. F-ratio values were generated using the six replicates of class 1 and class 2. Analytes are listed based on their order in Table E.1. Both the hit numbers and values (hit # / value) are provided.

Compound	F-ratio Analysis	1v1 Analysis					
		List #1	List #2	List #3	List #4	List #5	List #6
2-methylthiophene	#4 / 9.24×10^4	#4 / 35.5 %	#4 / 35.3 %	#4 / 35.5 %	#4 / 35.4 %	#4 / 35.5 %	#4 / 35.5 %
α-pinene	#8 / 6.01×10^4	#11 / 23.9 %	#11 / 23.7 %	#11 / 24.5 %	#12 / 24.1 %	#11 / 24.1 %	#11 / 24.2 %
1-bromooctane	#9 / 5.94×10^4	#3 / 35.6 %	#3 / 36.3 %	#3 / 36.6 %	#3 / 36.1 %	#3 / 36.4 %	#3 / 36.5 %
1,6-dichlorohexane	#6 / 6.43×10^4	#6 / 31.2 %	#7 / 31.2 %	#6 / 31.6 %	#7 / 31.2 %	#6 / 31.6 %	#6 / 31.6 %
phenylethyl alcohol	#18 / 1.08×10^4	#10 / 25.6 %	#10 / 25.4 %	#10 / 25.8 %	#10 / 25.8 %	#10 / 26 %	#10 / 26 %
2-dodecanone	#20 / 8.20×10^3	#5 / 34.5 %	#5 / 34.7 %	#5 / 34.9 %	#5 / 34.7 %	#5 / 34.8 %	#5 / 35.2 %
2-heptanone	#7 / 6.12×10^4	#16 / 17.4 %	#16 / 17.2 %	#16 / 17.4 %	#17 / 17.2 %	#16 / 16.9 %	#18 / 17.5 %
butyl sulfide	#1 / 1.77×10^5	#9 / 27.4 %	#9 / 27.9 %	#9 / 27.9 %	#9 / 28.2 %	#9 / 28.5 %	#9 / 29 %
methyl salicylate	#16 / 1.50×10^4	#15 / 17.9 %	#17 / 17 %	#15 / 17.9 %	#11 / 24.5 %	#15 / 18.2 %	#15 / 18.7 %
methyl decanoate	#12 / 3.48×10^4	#1 / 41.2 %	#1 / 41.8 %	#1 / 41.2 %	#1 / 41.7 %	#1 / 41.7 %	#1 / 41.7 %
citral	#15 / 1.66×10^4	#2 / 36.9 %	#2 / 37.7 %	#2 / 37.2 %	#2 / 37.8 %	#2 / 37.9 %	#2 / 38.3 %
dibenzylamine	#21 / 7.85×10^3	#8 / 30.3 %	#8 / 29.9 %	#8 / 30.8 %	#8 / 29.1 %	#8 / 31.3 %	#8 / 31.4 %
methyl caproate	#2 / 1.04×10^5	#14 / 18.6 %	#15 / 18.1 %	#13 / 18.8 %	#16 / 19.7 %	#14 / 18.6 %	#16 / 18.6 %
1-bromoheptane	#5 / 8.96×10^4	#19 / 9.7 %	#21 / 9.6 %	#21 / 9.7 %	#22 / 9.7 %	#21 / 9.9 %	#24 / 9.8 %
1-nonanol	#19 / 8.24×10^3	#18 / 10 %	#20 / 10.1 %	#20 / 10.6 %	#21 / 10.9 %	#20 / 10.8 %	#23 / 10.7 %
2-undecanone	#13 / 2.90×10^4	#12 / 20.9 %	#12 / 21.3 %	#12 / 20.8 %	#14 / 20.7 %	#12 / 21.9 %	#12 / 21.7 %
diphenyl sulfide	#10 / 5.92×10^4	#17 / 10.8 %	#19 / 11.9 %	#19 / 12.3 %	#19 / 12 %	#18 / 13.6 %	#21 / 13.9 %

dibutyl phthalate	#14 / 2.32×10^4	#7 / 31.1 %	#6 / 31.3 %	#7 / 31.3 %	#6 / 31.4 %	#7 / 31.6 %	#7 / 31.4 %

Table E.3. Jensen-Shannon divergence hit lists for the spiked diesel comparison. These hit lists were generated using paired replicate chromatograms from class 1 and class 2. Analytes are listed based on their order in Table E.1. Both the hit numbers and values (hit # / value) are provided. Analytes not discovered in the hit list are listed as NF / NA (Not Found/Not Applicable).

Compound	List #1	List #2	List #3	List #4	List #5	List #6
2-methylthiophene	# 24 / 0.63	# 30 / 0.63	# 35 / 0.63	# 25 / 0.63	# 17 / 0.63	# 22 / 0.63
α-pinene	# 54 / 0.42	# 62 / 0.41	# 77 / 0.26	# 51 / 0.42	# 49 / 0.4	# 46 / 0.42
1-bromooctane	# 39 / 0.5	# 46 / 0.5	# 43 / 0.52	# 35 / 0.51	# 34 / 0.51	# 34 / 0.51
1,6-dichlorohexane	# 27 / 0.61	# 31 / 0.61	# 30 / 0.7	# 30 / 0.61	# 23 / 0.61	# 25 / 0.62
phenylethyl alcohol	# 11 / 1.03	# 16 / 0.84	# 15 / 0.9	# 12 / 0.95	# 5 / 1.01	# 9 / 0.92
2-dodecanone	NF / NA	NF / NA	NF / NA	NF / NA	NF / NA	NF / NA
2-heptanone	# 19 / 0.75	# 20 / 0.75	# 24 / 0.76	# 17 / 0.75	# 11 / 0.75	# 15 / 0.75
butyl sulfide	NF / NA	NF / NA	# 211 / 0.1	NF / NA	NF / NA	NF / NA
methyl salicylate	# 23 / 0.65	# 28 / 0.66	# 32 / 0.65	# 23 / 0.67	# 16 / 0.65	# 19 / 0.67
methyl decanoate	# 33 / 0.55	# 44 / 0.51	# 45 / 0.47	# 24 / 0.64	# 35 / 0.5	# 33 / 0.51
citral	NF / NA	NF / NA	NF / NA	NF / NA	# 328 / 0.02	# 319 / 0.02
dibenzylamine	# 268 / 0.06	# 236 / 0.09	# 243 / 0.09	# 289 / 0.05	# 201 / 0.09	# 214 / 0.09
methyl caproate	NF / NA	NF / NA	NF / NA	NF / NA	# NF / NA	NF / NA
1-bromoheptane	# 223 / 0.1	NF / NA	# 193 / 0.12	# 239 / 0.08	# 208 / 0.08	NF / NA
1-nonanol	# 259 / 0.07	# 227 / 0.09	# 270 / 0.07	# 229 / 0.09	# 174 / 0.11	# 175 / 0.12
2-undecanone	# 29 / 0.58	# 36 / 0.58	# 46 / 0.48	# 38 / 0.49	# 37 / 0.49	# 36 / 0.5
diphenyl sulfide	NF / NA	NF / NA	NF / NA	NF / NA	NF / NA	NF / NA
dibutyl phthalate	# 63 / 0.37	# 72 / 0.32	# 65 / 0.33	# 76 / 0.26	# 66 / 0.3	# 67 / 0.29

Table E.4. Hit lists for the spiked diesel comparison generated by calculating the absolute difference between the paired replicate total ion current (TIC) chromatograms from class 1 and class 2. Analytes are listed based on their order in Table E.1. Both the hit numbers and values (hit # / value) are provided. Analytes not discovered in the hit list are listed as NF / NA (Not Found/Not Applicable).

Compound	List #1	List #2	List #3	List #4	List #5	List #6
2-methylthiophene	# 15 / 7.07	# 22 / 7.34	# 23 / 7.14	# 60 / 6.97	# 23 / 7.19	# 23 / 6.92
α-pinene	# 116 / 3.19	# 116 / 3.3	# 147 / 2.88	# 238 / 3.46	# 208 / 2.52	# 134 / 3.15
1-bromooctane	# 53 / 4.79	# 53 / 4.81	# 41 / 5.25	# 140 / 4.96	# 47 / 4.96	# 59 / 4.97
1,6-dichlorohexane	# 18 / 6.87	# 18 / 8.45	# 26 / 6.66	# 70 / 6.53	# 28 / 6.84	# 26 / 6.67
phenylethyl alcohol	# 256 / 1.93	# 256 / 2.13	# 253 / 1.91	# 487 / 2	# 286 / 2	# 250 / 2.01
2-dodecanone	# 156 / 2.65	# 156 / 2.84	# 133 / 3.1	# 673 / 1.18	# 496 / 1.13	# 532 / 0.98
2-heptanone	# 14 / 7.22	# 14 / 8.94	# 10 / 9.57	# 225 / 3.57	# 63 / 4.6	# 65 / 4.78
butyl sulfide	# 7 / 8.87	# 7 / 11.47	# 11 / 9.51	# 29 / 8.47	# 14 / 8.87	# 10 / 9.02
methyl salicylate	# 13 / 7.28	# 13 / 8.94	# 22 / 7.22	# 59 / 7.04	# 17 / 7.89	# 17 / 7.4
methyl decanoate	# 67 / 4.14	# 67 / 4.35	# 157 / 2.74	# 82 / 6.23	# 222 / 2.42	# 133 / 3.17
citral	# 222 / 2.21	# 231 / 2.31	# 244 / 1.97	# 463 / 2.02	# 255 / 2.17	# 220 / 2.22
dibenzylamine	# 322 / 1.59	# 322 / 1.9	# 311 / 1.42	# 592 / 1.83	# 453 / 1.34	# 266 / 1.94
methyl caproate	# 2 / 18.68	# 5 / 11.77	# 2 / 15.45	# 33 / 8.25	# 7 / 11.93	# 3 / 13.49
1-bromoheptane	# 4 / 13.9	# 4 / 12.44	# 4 / 11.15	# 9 / 12.11	# 5 / 13.3	# 2 / 14.66
1-nonanol	# 121 / 3.08	# 121 / 3.22	# 99 / 3.57	# 207 / 3.76	# 43 / 5.12	# 54 / 5.21
2-undecanone	# 46 / 5.1	# 46 / 5.08	# 42 / 5.25	# 83 / 6.17	# 45 / 5.08	# 36 / 5.54
diphenyl sulfide	# 1 / 21.47	# 1 / 20.63	# 1 / 21.24	# 2 / 19.74	# 2 / 22.66	# 1 / 22.74
dibutyl phthalate	# 22 / 6.68	# 23 / 7.19	# 17 / 7.64	# 48 / 7.42	# 18 / 7.85	# 13 / 7.84

Table E.5. Hit lists for the spiked diesel comparison generated by calculating the absolute difference on every *m/z* between the paired chromatograms from class 1 and class 2. Analytes are listed based on their order in Table E.1. Both the hit numbers and values (hit # / value) are provided. Analytes not discovered in the hit list are listed as NF / NA (Not Found/Not Applicable).

Compound	List #1	List #2	List #3	List #4	List #5	List #6
2-methylthiophene	# 1 / 0.405	# 1 / 0.500	# 1 / 0.407	# 1 / 0.393	# 1 / 0.410	# 1 / 0.392
α-pinene	# 249 / 0.005	# 313 / 0.005	# 387 / 0.002	# 471 / 0.007	# 513 / 0.002	# 289 / 0.004
1-bromooctane	# 49 / 0.016	# 53 / 0.018	# 48 / 0.018	# 140 / 0.016	# 55 / 0.018	# 52 / 0.019

1,6-dichlorohexane	# 12 / 0.037	# 18 / 0.037	# 14 / 0.036	# 18 / 0.037	# 20 / 0.038	# 16 / 0.038
phenylethyl alcohol	# 431 / 0.002	# 386 / 0.004	# 318 / 0.003	# 647 / 0.003	# 310 / 0.004	# 349 / 0.003
2-dodecanone	# 122 / 0.010	# 131 / 0.011	# 145 / 0.009	# 399 / 0.008	# 203 / 0.008	# 181 / 0.008
2-heptanone	# 21 / 0.028	# 14 / 0.046	# 11 / 0.043	# 23 / 0.032	# 26 / 0.031	# 21 / 0.037
butyl sulfide	# 3 / 0.191	# 3 / 0.193	# 3 / 0.200	# 3 / 0.185	# 3 / 0.189	# 3 / 0.198
methyl salicylate	# 33 / 0.020	# 49 / 0.018	# 42 / 0.019	# 73 / 0.020	# 44 / 0.021	# 50 / 0.019
methyl decanoate	# 13 / 0.037	# 24 / 0.029	# 16 / 0.032	# 26 / 0.031	# 23 / 0.034	# 28 / 0.030
citral	# 23 / 0.024	# 35 / 0.022	# 20 / 0.026	# 48 / 0.023	# 40 / 0.029	# 40 / 0.022
dibenzylamine	# 433 / 0.002	# 694 / 0.001	# 497 / 0.001	# 659 / 0.003	# 457 / 0.002	# 405 / 0.002
methyl caproate	# 2 / 0.315	# 2 / 0.315	# 2 / 0.286	# 2 / 0.241	# 2 / 0.276	# 2 / 0.268
1-bromoheptane	# 10 / 0.039	# 19 / 0.037	# 12 / 0.042	# 24 / 0.032	# 19 / 0.038	# 12 / 0.042
1-nonanol	# 7 / 0.075	# 6 / 0.083	# 5 / 0.084	# 8 / 0.082	# 5 / 0.086	# 6 / 0.087
2-undecanone	# 6 / 0.078	# 7 / 0.079	# 6 / 0.080	# 7 / 0.083	# 6 / 0.082	# 7 / 0.082
diphenyl sulfide	# 4 / 0.154	# 4 / 0.142	# 4 / 0.140	# 5 / 0.139	# 4 / 0.161	# 4 / 0.157
dibutyl phthalate	# 149 / 0.008	# 251 / 0.007	# 174 / 0.007	# 467 / 0.007	# 232 / 0.007	# 166 / 0.009

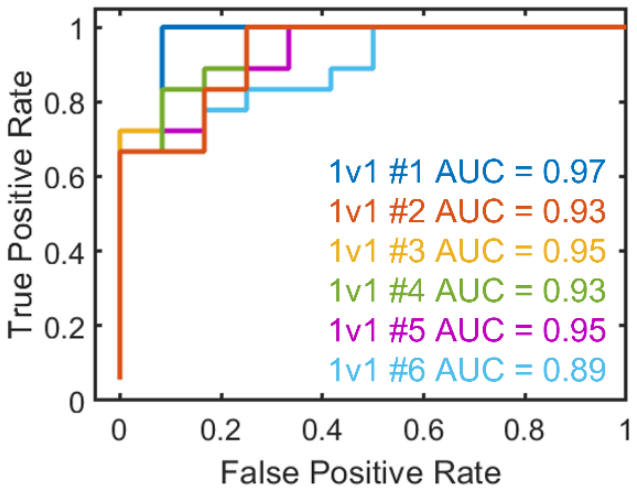


Figure E.2. Receiver operating characteristic (ROC) curves for the six hit lists generated using 1v1 analysis. ROC curves were generated by analyzing the top 30 hits in each list and the respective area under the curve (AUC) is also provided. The minor differences between the ROC curves developed for each hit list are due to the differences in ranking for each analyte and the number of false positives in the top 30 hits. For example. Hit list #6 for the 1v1 analyses is the only ROC curve to have an AUC less than 0.93 since more redundant hits were interspersed among the true positive hits (Table E.2). These comparable AUCs implies that tile-based 1v1 analysis has a similar performance in discovering analytes between the replicate hit lists.

Table E.6. Comparison of the average match values (MV) calculated for the 18 spiked analytes using the initial hit spectrum, MCR-ALS (all *m/z* and filtered *m/z*), PARAFAC, CCE-MSP (F-ratio and 1v1 Analyses), and CCE-MSP assisted MCR-ALS.[a]

Compound	$R_{s,2D}$	s_{int}/s_A	MV Initial	MV Standard MCR-ALS (All *m/z*)	MV PARAFAC	MV CCE-MSP F-ratio	MV CCE-MSP 1v1	MV CCE-MSP Assisted MCR-ALS	MV Standard MCR-ALS (Filtered *m/z*)
2-methylthiophene	1.25	0.70	914 ± 1	911 ± 14	980 ± 4	985 ± 1	986 ± 2	994 ± 1	990 ± 1
α-pinene	0.42	1.12	812 ± 4	873 ± 1	854 ± 2	988 ± 18	977 ± 16	997 ± 1	992 ± 1
1-bromooctane	0.73	0.26	737 ± 7	800 ± 2	813 ± 6	987 ± 12	986 ± 1	986 ± 1	982 ± 1
1,6-dichlorohexane	0.90	0.38	841 ± 7	828 ± 25	806 ± 49	988 ± 2	989 ± 2	964 ± 41	912 ± 1
phenylethyl alcohol	1.15	1.03	697 ± 9	832 ± 58	974 ± 1	792 ± 3	813 ± 19	996 ± 2	995 ± 2
2-dodecanone	0.30	24.51	147 ± 2	173 ± 1	170 ± 3	796 ± 15	753 ± 32	985 ± 5	196 ± 1
2-heptanone	0.42	32.24	149 ± 5	342 ± 37	625 ± 14	943 ± 10	940 ± 22	999 ± 1	455 ± 1
butyl sulfide	0.67	1.61	858 ± 4	909 ± 1	858 ± 10	993 ± 1	993 ± 1	996 ± 1	993 ± 1
methyl salicylate	0.41	0.33	754 ± 17	852 ± 3	880 ± 3	997 ± 3	996 ± 1	994 ± 1	985 ± 3
methyl decanoate	0.34	29.80	159 ± 1	131 ± 1	128 ± 1	919 ± 15	908 ± 39	964 ± 3	605 ± 1
citral	0.37	0.81	495 ± 3	709 ± 9	317 ± 5	753 ± 10	816 ± 31	916 ± 19	870 ± 12
dibenzylamine	0.82	3.84	337 ± 19	191 ± 1	195 ± 13	936 ± 4	918 ± 19	840 ± 10	358 ± 5
methyl caproate	0.39	21.16	300 ± 5	837 ± 7	945 ± 4	968 ± 5	982 ± 7	784 ± 9	783 ± 10
1-bromoheptane	0.96	7.72	925 ± 9	990 ± 1	984 ± 6	997 ± 1	996 ± 2	992 ± 1	993 ± 1
1-nonanol	0.38	12.78	181 ± 3	697 ± 37	358 ± 74	986 ± 2	987 ± 3	988 ± 2	986 ± 1
2-undecanone	0.52	14.79	359 ± 3	590 ± 5	749 ± 36	940 ± 9	980 ± 15	925 ± 3	872 ± 2
diphenyl sulfide	0.45	12.37	988 ± 2	986 ± 1	971 ± 7	993 ± 2	993 ± 1	991 ± 1	991 ± 1

dibutyl phthalate	0.33	28.62	153 ± 9	667 ± 2	783 ± 13	891 ± 9	879 ± 30	995 ± 1	995 ± 1
Total # of Hits with MV > 800			6	10	10	15	17	17	13

[a] The 2D chromatographic resolution ($R_{s,2D}$) and interference-to-analyte (s_{int}/s_A) was calculated using the lower spiked class data.

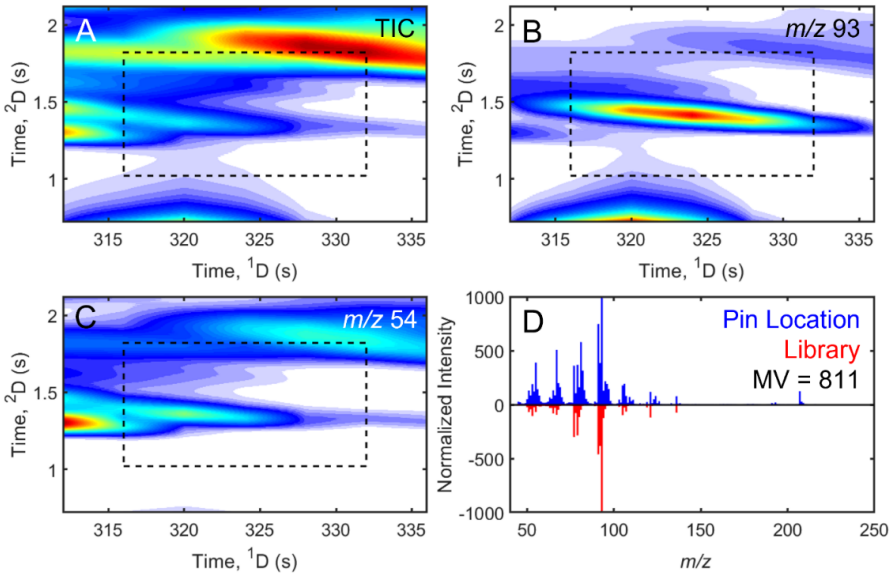


Figure E.3. Illustration of the surrounding chromatographic environment and mass spectrum for α-pinene. The 2D chromatogram of the region around the analyte is shown using the TIC (A), the top *m/z* discovered using 1v1 analysis (B), and a *m/z* selective for an interferent peak (C). The black dashed box represents the chosen tile size of 4 modulations on ^{1}D and 800 ms on ^{2}D. A comparison between the hit (blue) and library (red) spectra is also shown (D).

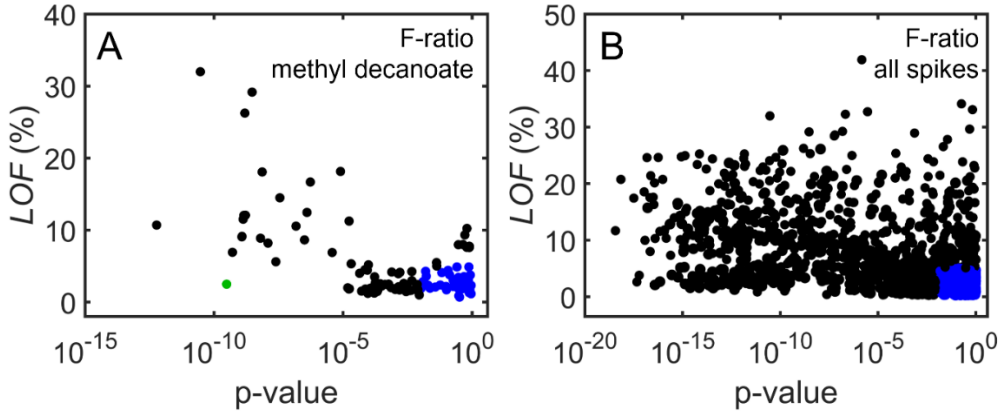


Figure E.4. Illustration of the signal consistency metrics for F-ratio analysis for methyl decanoate (A) and all spiked analytes (B). Pure interferent *m/z* (blue) were identified as *m/z* having a *LOF* ≤ 5 % and a p-value ≥ 0.01. The pure analyte *m/z* for methyl decanoate is shown in green.

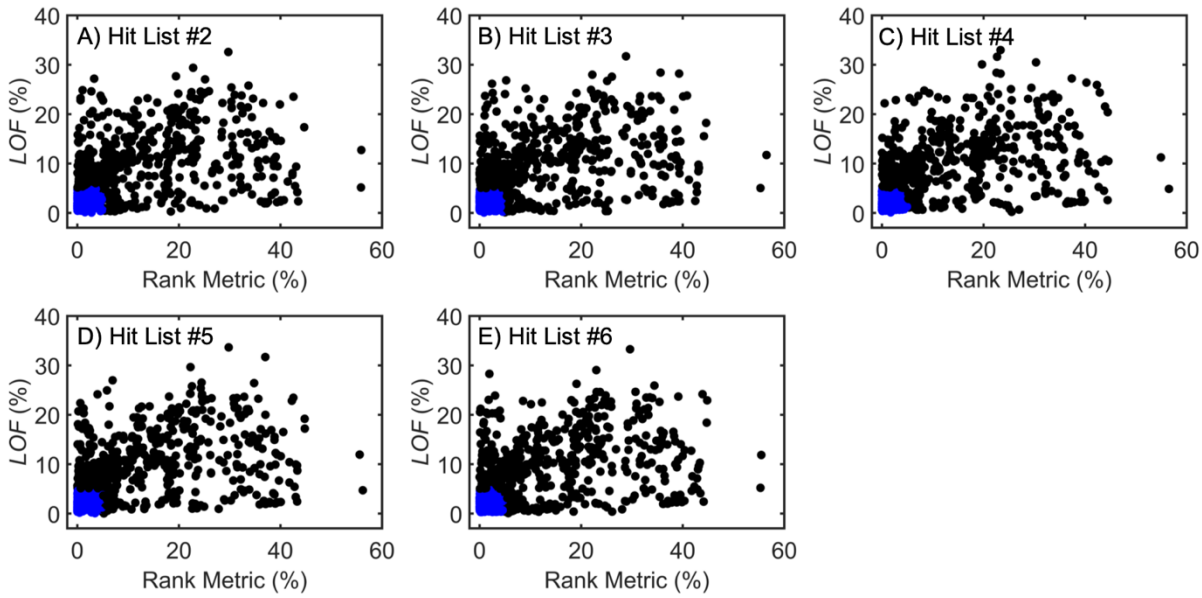


Figure E.5. Plots of *LOF* versus *RM* for all 18 spiked analytes discovered with the other five 1v1 comparisons. Pure interferent *m/z* (blue) were identified as *m/z* having a $LOF \leq 5$ % and a $RM \leq 5$ %. The overall shape of the plots shown here is like the plot in Figure 7.4B.

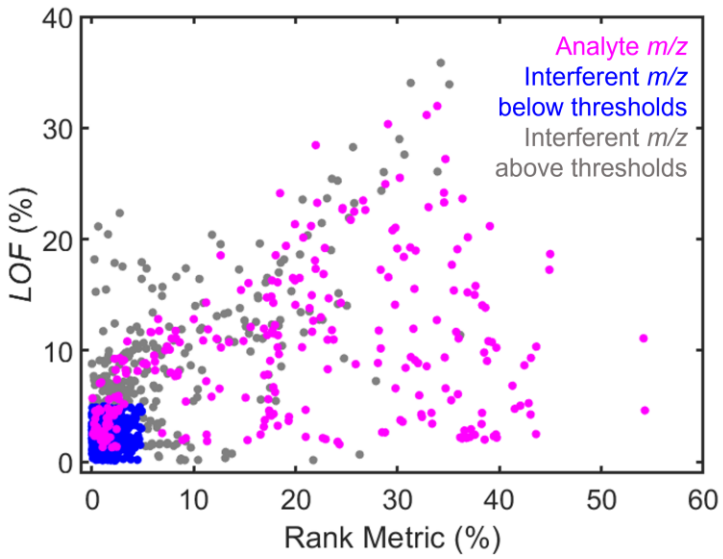


Figure E.6. Differentiation between analyte and interferent *m/z* based on their *LOF* and *RM*. The plot shown here uses the first 1v1 comparison. The *m/z* (pink) belongs to the analytes (with signal ≥ 1 % of the base peak in the library spectrum) but can also have contributions from the interferent peaks. Indeed, the pink *m/z* with *LOF*s and *RM*s ≤ 5 % are heavily dominated by the interferent peaks. Interferent *m/z* highlighted in blue had *LOF*s and *RM*s ≤ 5 % while interferent *m/z* shown in gray had *LOF*s and *RM*s ≥ 5 %. In total, 469 analyte *m/z* (pink) and 786 interferent *m/z* (blue and gray) were discovered. Within the *LOF* and *RM* threshold region, there are 191 interferent *m/z* and 39 pink *m/z* (analyte marginally contributing) after applying the *S/N* filter.

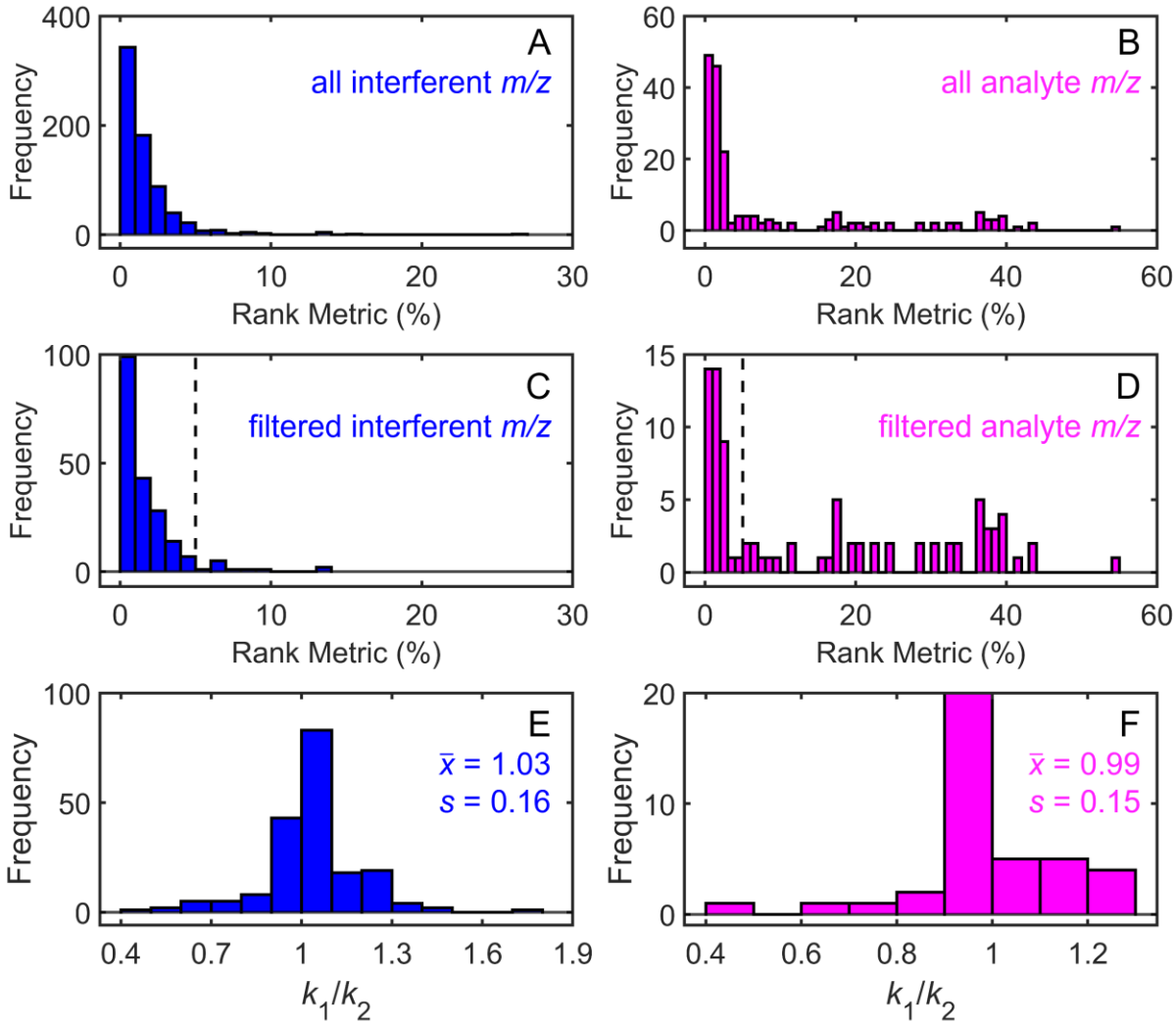


Figure E.7. Histograms demonstrating the selection of the *LOF* and *RM* filters for the interferent (blue) and potential analyte contributing (pink) *m/z* discovered. (A and B) Histograms of the *RM* for *m/z* with a *LOF* ≤ 5 %. There was a total of 704 interferent (A) and 186 analyte (B) *m/z* with *LOF*s ≤ 5 %. (C and D) Histograms of the *RM* after applying a 5 % signal threshold to remove *m/z* with a low *S/N*. There was a total of 202 interferent (C) and 90 analyte (D) *m/z* remaining after the *S/N* filter. The black dashed line represents the *RM* threshold of 5 %. A total of 191 interferent (C) and 39 analyte (D) *m/z* had a *RM* ≤ 5 %. (E and F) Histograms of the normalization factors, k_1/k_2, calculated for the remaining interferent and analyte *m/z (*observed in C and D) with a *LOF* and *RM* ≤ 5 %. The mean ($\bar{x}$) and standard deviation (*s*) for both distributions are provided. A *t*-test found that there was not a significant difference between the distributions for the interferent (E) and analyte (F) normalization factors ($p = 0.09$).

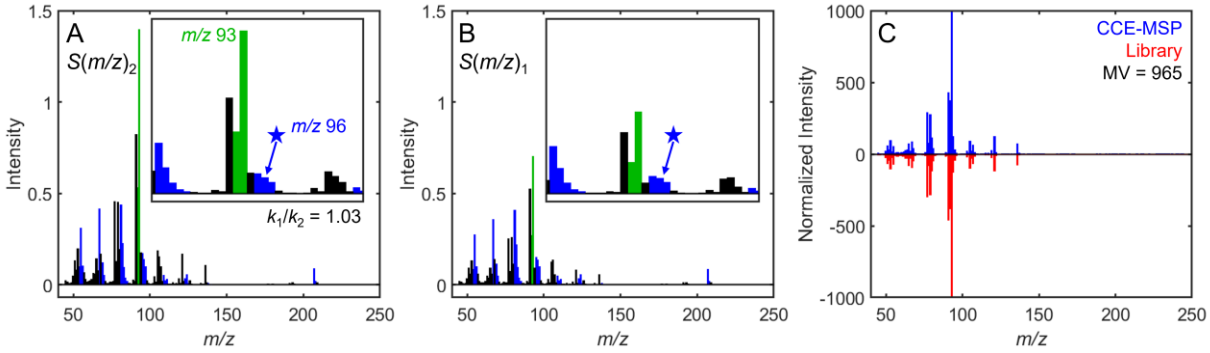


Figure E.8. Application of CCE-MSP to α-pinene using information from the 1v1 analysis. The hit spectrum in class 2, $S(m/z)_2$ (A), and in class 1, $S(m/z)_1$ (B), are shown. The *m/z* shown in A and B are colored according to designation as pure interferent *m/z* (blue), pure analyte *m/z* (green), and all other *m/z* (black). $S(m/z)_2$ is normalized by k_1/k_2, which equates to the signal ratio for a pure interferent *m/z* (indicated by the blue star). Insets: Zoom-in from 80-110 *m/z* to illustrate the pure interferent and analyte *m/z*. The scale for the y-axes of the insets is -0.07 to 1.5. (C) The two hit spectra are then subtracted from one another to generate the purified analyte spectrum (blue), which is compared against the library spectrum (red). A match value (MV) is also provided.

Table E.7. The MV acquired for the 18 spiked analytes with MCR-ALS and MCR-BANDS.[a]

Compound	Standard MCR-ALS (all *m/z*)	MCR-BANDS (all *m/z*)
2-methylthiophene	896	896-897
α-pinene	873	873
1-bromooctane	797	797
1,6-dichlorohexane	817	817
phenylethyl alcohol	891	797-905
2-dodecanone	171	171
2-heptanone	352	242-330
butyl sulfide	908	908
methyl salicylate	848	848
methyl decanoate	134	134
citral	712	712
dibenzylamine	191	191
methyl caproate	834	834
1-bromoheptane	990	987-990
1-nonanol	703	691-703
2-undecanone	584	584
diphenyl sulfide	986	986
dibutyl phthalate	666	666

[a] MCR-ALS and MCR-BANDS was performed using the first spiked diesel pairwise comparison. The same constraints were utilized for both algorithms.

MCR-BANDS determines if rotational ambiguities are present for an analyte by finding the minimum and maximum relative contribution of each component. We refer the reader to Jaumot and Tauler for additional information on the MCR-BANDS algorithm [5]. If rotational ambiguities are present, then the minimum and maximum relative contributions would not be equal, causing minor differences in the chromatographic peak profile and mass spectrum obtained. Hence, the MV obtained for the analyte with MCR-BANDS would be a range with the standard MCR-ALS result falling in that range. Conversely, if rotational ambiguities are not present, then the MV obtained with MCR-BANDS would not be a range and would equal the standard MCR-ALS result. Table E.7 shows that rotational ambiguities were only present for 5 of the 18 spiked analytes to a minor degree. Therefore, the poor performance of MCR-ALS to resolve many of these analytes was not due to this type of uncertainty.

[5] J. Jaumot, R. Tauler, MCR-BANDS: A user friendly MATLAB program for the evaluation of rotation ambiguities in Multivariate Curve Resolution, Chemom. Intell. Lab. Syst. 103 (2010) 96–107. https://doi.org/10.1016/j.chemolab.2010.05.020.

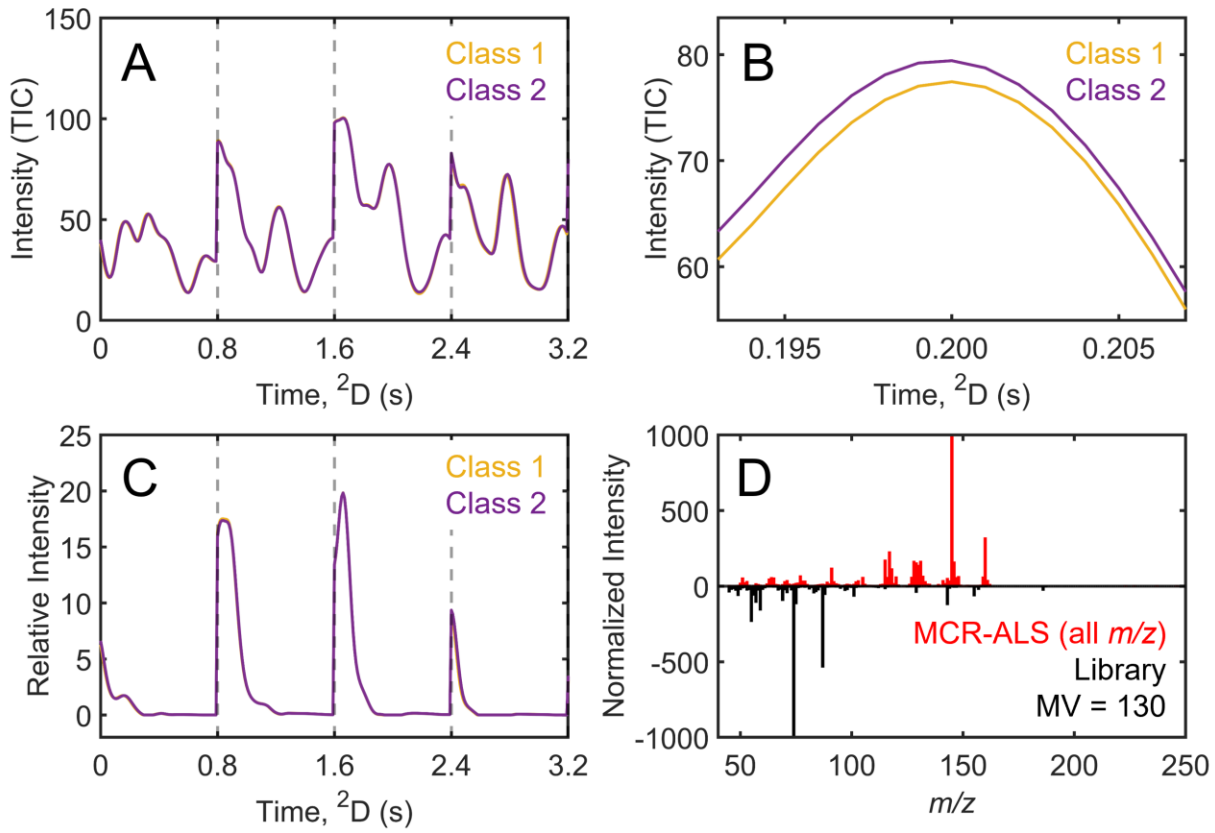


Figure E.9. Demonstration of MCR-ALS on methyl decanoate using all *m/z*. (A) The unfolded TIC chromatogram for the tile surrounding methyl decanoate. The gray dashed vertical lines represent each modulation in the tile. Chromatograms for class 1 and 2 are shown in yellow and purple, respectively. (B) Zoom-in of the TIC chromatogram in A to demonstrate that there is some discernable difference in signal between the two classes, which correlates to methyl decanoate. (C) The resolved retention profiles for the MCR-ALS component that had the highest MV to the library spectrum. (D) Comparison of the resolved mass spectrum for the MCR-ALS component in C (red) to the library spectrum of methyl decanoate (black).

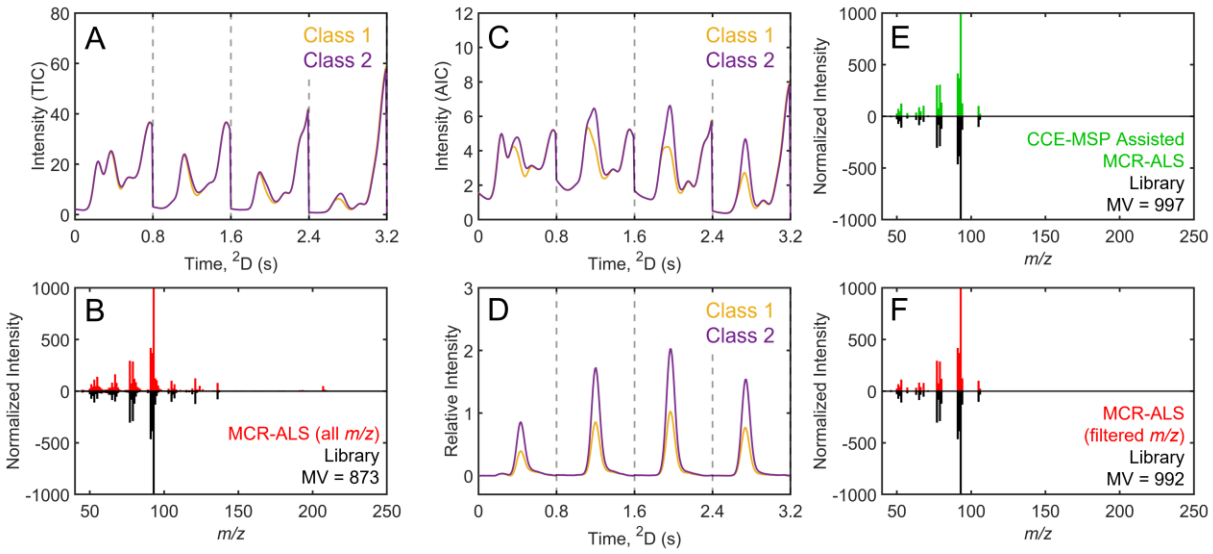


Figure E.10. Demonstration of CCE-MSP assisted MCR-ALS on α-pinene using the information from 1v1 analysis. (A) The unfolded TIC chromatogram for the tile surrounding methyl decanoate. The gray dashed vertical lines represent each modulation in the tile. Chromatograms for class 1 and 2 are shown in yellow and purple, respectively. (B) Reflection plot of the resolved mass spectrum for the MCR-ALS component that had the highest MV (red) and the library spectrum of methyl decanoate (black). (C) The unfolded analytical ion current (AIC) chromatogram created by summing together the 21 *m/z* above the *RM* and *LOF* thresholds. (D) The resolved retention profiles for the MCR-ALS component that had the highest MV to the library spectrum. (E) Comparison of the resolved mass spectrum for the MCR-ALS component in D (green) to the library spectrum of methyl decanoate (black). (F) Reflection plot of the initial MCR-ALS spectrum in B filtered down to the 21 *m/z* of interest (red) and the filtered library spectrum (black).

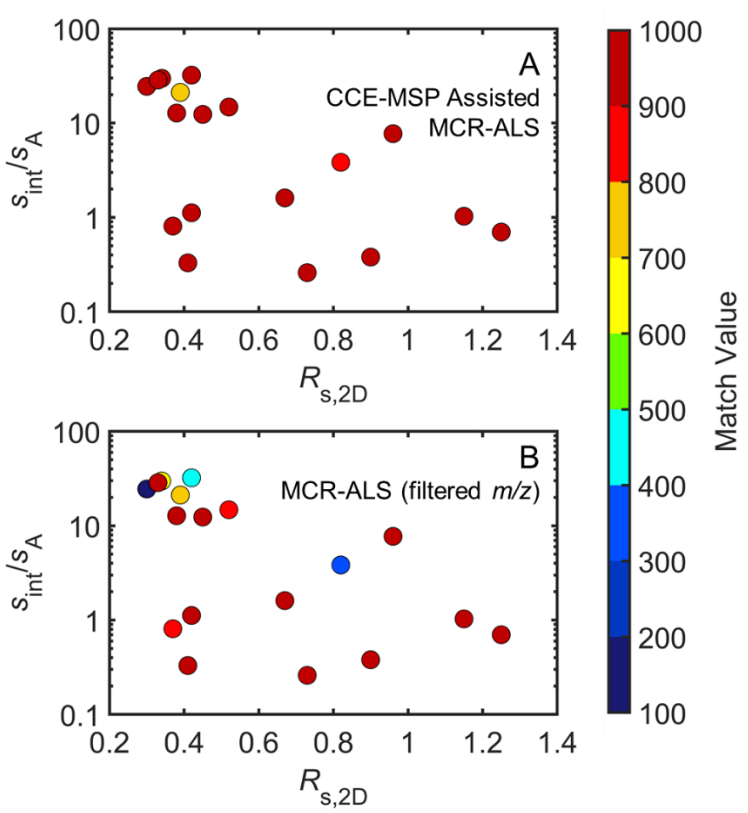


Figure E.11. Plots of the interference-to-analyte ratios (s_{int}/s_A) versus 2D resolution ($R_{s,2D}$) for the 18 spiked analytes. Each data point is colored according to the average MV determined with CCE-MSP assisted MCR-ALS (A) or the initial MCR-ALS using the filtered *m/z* (B).

Table E.8. Hit list for the first 1v1 comparison between the unmolded and molded cacao beans, shown in Figure 7.8. The largest *RM* and corresponding *m/z* are provided. For each hit and *m/z*, the peak area in the unmolded sample was divided by the peak area in the molded sample to produce a signal ratio ([Unmolded]/[Molded] = [U]/[M]), i.e., an apparent concentration ratio assuming a pure *m/z* is used. Each hit shaded in green was also discovered in the top 30 hits for F-ratio analysis (Table E.9).

Hit #	^1t_R (min)	^2t_R (s)	*RM*	*m/z*	Analyte	MV	[U]/[M]
1	6.94	0.49	99.8	45	2,3-Butanediol	910	933.06
2	7.19	1.35	99.3	43	Butanoic acid	829	291.31
3	2.37	0.89	99.3	45	Acetic acid	934	270.86
4	13.59	0.82	99.2	57	1-Hexanol, 2-ethyl-	929	247.52
5	6.22	1.49	98.9	73	Propanoic acid, 2-methyl-	813	177.11
6	7.22	0.31	97.3	45	2,3-Butanediol, [*S*-(*R**,*R**)]-	911	73.42
7	15.19	1.24	96.6	57	Nonanal	901	57.77
8	13.47	0.61	95.5	106	4-Cyanocyclohexene	878	45.46
9	19.32	0.43	94.5	71	Propanoic acid, 2-methyl-, 3-hydroxy-2,4,4-trimethylpentyl ester	832	35.29

10	12.44	1	92.8	42	Hexanoic acid	938	26.62
11	4.19	0.61	92.5	43	Heptane	921	25.77
12	18.52	0.17	91.4	85	Tridecane	947	22.23
13	4.67	0.49	91.1	45	Acetoin	806	0.05
14	19.42	0.05	91.0	85	Hexadecane	935	21.25
15	19.17	0.08	90.2	57	Tridecane, 2-methyl-	852	19.45
16	14.89	0.73	88.9	136	Pyrazine, tetramethyl-	931	0.06
17	19.24	0.16	87.1	57	Dodecane, 2,6,10-trimethyl-	932	14.45
18	20.67	0.23	86.7	43	Propanoic acid, 2-methyl-, 1-(1,1-dimethylethyl)-2-methyl-1,3-propanediyl ester	854	14.06
19	18.19	0.31	85.4	57	Octane, 2,3,7-trimethyl-	922	12.70
20	18.09	0.26	85.0	57	Dodecane, 2-methyl-	889	12.36
21	18.97	0.13	84.5	56	1-Decanol, 2-hexyl-	862	11.91
22	12.09	0.24	84.2	51	Benzaldehyde	932	11.66
23	14.52	0.2	83.6	51	Acetophenone	962	11.19
24	17.02	0.39	82.9	43	Dodecane	896	10.72
25	19.52	0.3	81.0	57	Dodecanal	921	9.57
26	19.17	0.5	80.7	43	Propanoic acid, 2-methyl-, 2,2-dimethyl-1-(2-hydroxy-1-methylethyl	835	9.39
27	18.29	0.23	80.6	57	Heptadecane, 2,6,10,14-tetramethyl	870	9.32
28	13.64	0.53	80.3	41	Cyclohexene, 1-methyl-4-(1-methylethenyl)-, (*S*)-	901	5.28
29	18.64	0.53	79.9	41	Undecanal	916	10.75
30	13.02	0.9	79.7	122	Pyrazine, trimethyl-	914	0.11
31	19.94	0.04	79.6	83	β-Patchoulene	845	8.81
32	10.97	1.18	79.4	67	Pyrazine, 2,3-dimethyl-	916	0.11
33	17.27	0.41	78.0	57	Undecane, 2,6-dimethyl-	907	8.09
34	17.94	0.3	77.7	41	2,3-Dimethyldecane	865	7.21
35	17.77	0.34	76.4	43	Undecane, 5-ethyl-	854	7.47
36	17.82	0.39	75.3	83	Cyclohexane, hexyl-	877	7.11
37	21.22	0.13	74.3	55	Benzoic acid, 2-ethylhexyl ester	833	6.79
38	21.72	0.24	74.3	43	2-Pentadecanone, 6,10,14-trimethyl-	806	6.76
39	21.37	1.5	73.8	55	*n*-Nonadecanol-1	804	6.64
40	8.32	0.76	73.1	67	1,3-Octadiene	903	0.11
41	8.54	0.07	72.8	71	1,6-Heptadien-4-ol	802	0.16
42	21.52	0.01	72.5	57	1-Hexadecanol, 2-methyl-	830	6.26
43	21.12	1.48	72.4	41	Oxalic acid, allyl octadecyl ester	823	5.92
44	13.74	1.08	72.3	78	Benzyl Alcohol	880	10.22
45	17.89	0.71	72.2	45	2-Propanol, 1-(2-butoxy-1-methylethoxy)-	809	5.95
46	14.87	0.99	72.1	43	Benzenemethanol, α,α-dimethyl-	911	6.17
47	15.47	1.08	71.7	65	Phenylethyl Alcohol	931	6.06
48	18.79	0.36	71.3	41	*n*-Butyric acid 2-ethylhexyl ester	884	6.43
49	1.64	0.72	71.2	45	Nitrous oxide	952	0.14
50	12.44	0.82	70.5	57	1-Octen-3-ol	908	0.17
51	8.87	1.35	70.1	56	2-Pentanone, 4-hydroxy-4-methyl-	866	6.32

52	1.82	0.49	69.8	43	Acetic acid, methyl ester	883	5.63
53	1.77	0.71	69.6	44	Carbon dioxide	961	0.18
54	17.17	1.14	68.9	83	Decanal	879	8.11
55	8.92	1.21	68.2	56	Butanoic acid, 3-methyl-	888	5.89
56	10.34	0.17	67.5	71	2,3-Octanedione	802	5.15
57	13.99	0.48	66.5	41	Hexane, 1-nitro-	915	6.33
58	16.07	1.01	64.4	77	Benzene, 1,2-dimethoxy-	800	4.62
59	14.92	1.4	62.7	43	2-Nonanone	812	4.37
60	9.87	1.35	60.1	43	1-Butanol, 3-methyl-, acetate	892	4.21
61	16.84	0.46	59.5	43	Cyclohexane, 1-methyl-2-pentyl-	861	3.94
62	11.92	0.53	59.3	43	2-Heptenal, (*Z*)-	840	3.92
63	16.64	0.83	59.1	51	2H-Pyran-3-ol, 6-ethenyltetrahydro-2,2,6-trimethyl-	848	0.26
64	21.97	0.15	58.3	100	*n*-Morpholinomethyl-isopropyl-sulfide	129	0.26
65	16.49	0.75	57.9	55	1-Nonanol	886	3.74
66	3.22	0.1	57.5	41	Butanal, 3-methyl-	872	0.25
67	16.34	0.61	56.6	51	Pyrazine, 3,5-diethyl-2-methyl-	833	0.33
68	16.17	0.41	56.3	41	Octane, 4-ethyl-	849	3.58
69	14.99	0.47	56.1	43	1-Heptanol, 2-propyl-	810	3.55
70	15.64	0.47	53.0	43	1-Undecene, 4-methyl-	801	3.25
71	9.17	1.11	51.4	55	Butanoic acid, 2-methyl-	844	3.11
72	1.54	0.13	50.9	43	Acetaldehyde	895	3.08
73	15.12	0.8	49.9	55	1-Octene, 6-methyl	815	2.99
74	14.47	0.8	49.7	42	1-Octanol	844	2.97
75	16.97	0.87	49.4	57	Octanoic acid, ethyl ester	878	2.95
76	10.37	1.24	46.8	77	2-Propenoic acid, butyl ester	831	2.76
77	21.09	0.15	46.7	43	2-Dodecanone	822	2.76
78	12.09	0.74	45.6	105	Benzene, 1-ethyl-3-methyl-	928	2.68
79	14.72	0.15	43.9	44	Octane, 3,5-dimethyl-	832	0.39
80	3.39	0.79	43.2	78	Benzene	840	2.52
81	14.84	0.49	42.8	43	1-Octanol, 2-butyl-	802	2.46
82	2.99	0.84	41.3	42	Tetrahydrofuran	876	2.41
83	12.84	0.76	38.2	77	Benzene, 1,2,3-trimethyl-	923	2.23
84	15.04	0.41	37.3	71	Undecane	883	2.19
85	12.54	0.78	33.4	105	Benzene, 1-ethyl-4-methyl-	872	2.00
86	14.89	0.45	33.3	43	4-Dodecene, (*E*)-	878	2.00

Table E.9. Top 30 hits for the unmolded versus molded comparison using tile-based F-ratio analysis. The largest F-ratio and corresponding *m/z* are provided. For each hit and *m/z*, the average peak area in the unmolded sample was divided by the average peak area in the molded sample to produce a signal ratio ([Unmolded]/[Molded] = [U]/[M]).

Hit #	^1t_R (min)	^2t_R (s)	F-ratio	*m/z*	Analyte	MV	[U]/[M]
1	2.32	0.85	4760	46	Acetic acid	945	245.78
2	7.27	0.3	2161	60	2,3-Butanediol, [*S*-(*R*,R**)]-	922	28.03
3	8.92	1.21	1698	56	Butanoic acid, 3-methyl-	888	23.82
4	7.19	1.35	1307	43	Butanoic acid	829	222.11
5	8.87	1.35	1284	56	2-Pentanone, 4-hydroxy-4-methyl-	866	36.45
6	14.87	0.76	826	49	Pyrazine, tetramethyl-	931	0.15
7	14.87	1.06	576	50	Benzenemethanol, α,α-dimethyl-	802	0.89
8	6.22	1.49	547	73	Propanoic acid, 2-methyl-	813	68.75
9	16.34	0.61	512	51	Pyrazine, 3,5-diethyl-2-methyl-	833	0.23
10	13.47	0.61	485	106	4-Cyanocyclohexene	884	23.18
11	6.97	0.45	461	60	2,3-Butanediol	908	190.62
12	12.42	1.06	448	45	Hexanoic acid	808	61.60
13	2.12	0.01	421	44	Carbon dioxide	905	0.07
14	8.32	0.76	416	67	1,3-Octadiene	903	0.04
15	12.99	0.91	409	50	Pyrazine, trimethyl-	914	0.16
16	13.74	1.08	348	78	Benzyl Alcohol	880	20.93
17	16.97	0.87	248	57	Octanoic acid, ethyl ester	816	5.12
18	10.34	0.17	241	71	2,3-Octanedione	802	50.50
19	16.64	0.83	235	51	2H-Pyran-3-ol, 6-ethenyltetrahydro-2,2,6-trimethyl-	821	0.38
20	9.87	1.35	226	43	1-Butanol, 3-methyl-, acetate	892	76.73
21	15.19	1.2	213	68	Nonanal	912	23.64
22	17.17	1.14	213	83	Decanal	894	7.54
23	16.84	0.46	208	43	Cyclohexane, 1-methyl-2-pentyl-	861	3.52
24	1.64	0.72	178	45	Nitrous oxide	952	0.13
25	17.87	0.71	178	45	2-Propanol, 1-(2-butoxy-1-methylethoxy)-	816	5.87
26	13.57	0.84	170	95	1-Hexanol, 2-ethyl-	924	14.58
27	9.19	1.11	167	56	Butanoic acid, 2-methyl-	856	38.07
28	19.32	0.43	156	55	Propanoic acid, 2-methyl-, 3-hydroxy-2,4,4-trimethylpentyl ester	840	30.79
29	10.37	1.24	156	77	2-Propenoic acid, butyl ester	837	8.88
30	14.52	0.31	133	43	Acetophenone	810	54.40

Table E.10. Top 30 hits for the second 1v1 comparison between the unmolded and molded cacao beans. The largest *RM* and corresponding *m/z* are provided. For each hit and *m/z*, the peak area in the unmolded sample was divided by the peak area in the molded sample to produce a signal ratio ([Unmolded]/[Molded] = [U]/[M]).

Hit #	$^{1}t_{R}$ (min)	$^{2}t_{R}$ (s)	*RM*	*m/z*	Analyte	MV	[U]/[M]
1	2.34	0.86	99.4	43	Acetic acid	945	331.57
2	12.44	0.84	99.1	57	1-Octen-3-ol	927	0.01
3	1.17	0.31	94.6	44	Carbon dioxide	813	0.03
4	13.59	0.83	89.2	70	1-Hexanol, 2-ethyl-	924	17.58
5	15.19	1.24	88.6	41	Nonanal	891	20.63
6	12.99	0.9	85.3	42	Pyrazine, trimethyl-	916	0.08
7	14.87	0.76	82.4	54	Pyrazine, tetramethyl-	931	0.10
8	10.97	1.18	81.2	67	Pyrazine, 2,3-dimethyl-	918	0.10
9	1.47	0.02	80.7	44	Nitrous oxide	850	0.11
10	19.32	0.42	80.3	43	Propanoic acid, 2-methyl-, 3-hydroxy-2,4,4-trimethylpentyl ester	844	9.15
11	21.02	0.06	79.7	100	N-Morpholinomethyl-isopropyl-sulfide	835	0.11
12	4.19	0.61	78.9	43	Heptane	910	8.47
13	6.07	1.49	77.8	41	Propanoic acid, 2-methyl-	885	9.79
14	15.12	0.73	75.8	43	1,6-Octadien-3-ol, 3,7-dimethyl- / 1-Octene, 6-methyl-	831	7.21
15	2.49	0.62	74.0	43	3-Buten-2-ol, 2-methyl-	881	0.15
16	19.24	0.31	73.2	57	Dodecane, 2,6,10-trimethyl-	845	6.45
17	14.89	0.46	72.1	41	3-Undecene, (*E*)-	828	0.76
18	19.14	0.08	70.7	71	Tridecane, 2-methyl-	901	5.83
19	19.84	0.02	66.7	43	Decane, 2-methyl-	859	5.01
20	18.69	0.15	66.6	57	Tetradecane	867	4.99
21	19.02	0.1	65.7	43	Oxalic acid, isobutyl nonyl ester	854	4.83
22	19.84	0.26	64.5	43	5,9-Undecadien-2-one, 6,10-dimethyl-, (*E*)-	850	4.63
23	18.54	0.16	64.5	71	Tridecane	922	4.63
24	16.09	1	63.3	41	Benzene, 1,2-dimethoxy-	883	2.29
25	18.62	0.15	62.5	71	Decane	860	4.33
26	19.34	0.07	62.4	43	1-Docosene	872	4.25
27	18.97	0.11	61.9	50	1-Decanol, 2-hexyl-	844	4.18
28	12.09	0.24	61.4	57	Benzaldehyde	935	4.17
29	19.17	0.49	61.3	43	Propanoic acid, 2-methyl-, 2,2-dimethyl-1-(2-hydroxy-1-methylethyl)propyl ester	878	4.04
30	18.64	0.53	60.5	43	Undecanal	923	4.06

Table E.11. Top 30 hits for the third 1v1 comparison between the unmolded and molded cacao beans. The largest *RM* and corresponding *m/z* are provided. For each hit and *m/z*, the peak area in the unmolded sample was divided by the peak area in the molded sample to produce a signal ratio ([Unmolded]/[Molded] = [U]/[M]).

Hit #	$^{1}t_{R}$ (min)	$^{2}t_{R}$ (s)	*RM*	*m/z*	Analyte	MV	[U]/[M]
1	2.57	1.03	99.8	45	Acetic acid	934	1319.63
2	9.44	1.05	95.6	41	Butanoic acid, 2-methyl-	886	72.44
3	14.89	0.91	95.5	54	Pyrazine, tetramethyl-	889	0.02
4	4.52	0.49	95.1	45	Acetoin	805	39.98
5	1.14	0.5	89.2	44	Carbon dioxide	921	0.06
6	12.62	1.4	88.6	57	3-Octanone	874	0.06
7	14.52	0.2	87.9	51	Acetophenone	960	15.56
8	15.19	1.25	87.3	57	Nonanal	904	14.69
9	13.84	1.04	86.1	43	2-Heptanol, acetate	871	13.43
10	12.42	1	86.1	41	Hexanoic acid	931	9.45
11	13.59	0.82	85.9	43	1-Hexanol, 2-ethyl-	919	13.23
12	12.44	0.82	85.7	57	1-Octen-3-ol	911	0.08
13	15.09	0.78	80.8	45	2-Nonanol	883	0.11
14	1.52	1.44	80.7	44	Nitrous oxide	928	0.11
15	1.57	0.13	79.6	43	Acetaldehyde	942	8.81
16	19.32	0.43	79.2	71	Propanoic acid, 2-methyl-, 3-hydroxy-2,4,4-trimethylpentyl ester	841	8.59
17	12.07	0.25	77.6	51	Benzaldehyde	943	7.91
18	16.34	0.59	77.1	41	Pyrazine, 3,5-diethyl-2-methyl-	845	0.47
19	14.92	0.44	75.9	41	3-Undecene, (*E*)-	801	0.63
20	21.97	0.16	74.4	100	*n*-Morpholinomethyl-isopropyl-sulfide	866	0.15
21	18.09	0.6	71.1	57	Nonanoic acid	873	5.88
22	14.94	1.39	68.9	43	2-Nonanone	922	0.18
23	19.39	0.13	67.4	91	α-copaene	896	0.19
24	17.17	1.14	63.3	41	Decanal	896	5.66
25	16.07	1.01	60.9	41	Benzene, 1,2-dimethoxy-	832	2.50
26	15.47	1.08	58.6	91	Phenylethyl Alcohol	925	3.83
27	13.64	0.54	58.4	41	Cyclohexene, 1-methyl-4-(1-methylethenyl)-, (*S*)-	912	3.23
28	3.17	0.1	57.5	41	Butanal, 3-methyl-	881	0.24
29	13.99	0.48	56.5	43	Hexane, 1-nitro-	883	3.59
30	19.79	0.11	55.9	91	1-Decanol, 2-hexyl-	858	0.28

Table E.12. Top 30 hits for the fourth 1v1 comparison between the unmolded and molded cacao beans. The largest *RM* and corresponding *m/z* are provided. For each hit and *m/z*, the peak area in the unmolded sample was divided by the peak area in the molded sample to produce a signal ratio ([Unmolded]/[Molded] = [U]/[M]).

Hit #	$^{1}t_{R}$ (min)	$^{2}t_{R}$ (s)	*RM*	*m/z*	Analyte	MV	[U]/[M]
1	2.42	0.9	99.8	43	Acetic acid	937	960.81
2	1.44	1.43	97.5	44	Carbon dioxide	994	0.01
3	21.97	0.15	91.8	100	*n*-Morpholinomethyl-isopropyl-sulfide	854	0.04
4	12.44	1	91.6	42	Hexanoic acid	949	22.76
5	19.32	0.43	91.4	71	Propanoic acid, 2-methyl-, 3-hydroxy-2,4,4-trimethylpentyl ester	841	22.10
6	13.59	0.82	88.5	70	1-Hexanol, 2-ethyl-	924	16.38
7	4.17	0.61	88.3	43	Heptane	918	16.11
8	13.02	0.9	86.7	42	Pyrazine, trimethyl-	908	0.07
9	10.97	1.18	84.6	67	Pyrazine, 2,3-dimethyl-	914	0.08
10	14.89	0.72	83.5	136	Pyrazine, tetramethyl-	928	0.09
11	12.09	0.23	76.2	106	Benzaldehyde	943	7.39
12	14.94	0.48	76.1	41	3-Undecene, (*E*)-	825	0.57
13	8.52	0.08	75.2	43	1,6-Heptadien-4-ol	801	0.14
14	19.42	0.06	75.1	85	Hexadecane	937	7.03
15	12.44	0.82	74.7	57	1-Octen-3-ol	890	0.14
16	14.52	0.19	74.2	105	Acetophenone	961	6.76
17	19.17	0.09	73.3	57	Nonadecane	868	6.50
18	20.67	0.23	72.2	71	Propanoic acid, 2-methyl-, 1-(1,1-dimethylethyl)-2-methyl-1,3-propanediyl ester	863	6.19
19	18.52	0.18	72.1	85	Tridecane	937	6.18
20	5.37	0.06	71.9	41	1-Butanol, 3-methyl-	802	0.16
21	19.07	0.1	70.9	71	Tridecane, 4-methyl-	888	5.88
22	9.42	0.91	68.2	65	Ethylbenzene	948	0.19
23	19.94	0.09	68.2	91	β-Patchoulene	892	0.19
24	18.64	0.53	67.6	57	Undecanal	893	5.07
25	19.17	0.51	67.4	57	Propanoic acid, 2-methyl-, 2,2-dimethyl-1-(2-hydroxy-1-methylethyl)propyl ester	812	5.15
26	16.07	1.01	66.2	41	Benzene, 1,2-dimethoxy-	861	3.16
27	19.24	0.31	65.8	57	Dodecane, 2,6,10-trimethyl-	841	4.87
28	19.84	0.03	64.3	71	Heptadecane, 2,6-dimethyl-	874	4.60
29	18.09	0.26	63.9	43	Tridecane, 3-methyl-	893	4.53
30	10.32	0.92	63.7	91	*p*-Xylene	936	0.22

Table E.13. Top 30 hits for the fifth 1v1 comparison between the unmolded and molded cacao beans. The largest *RM* and corresponding *m/z* are provided. For each hit and *m/z*, the peak area in the unmolded sample was divided by the peak area in the molded sample to produce a signal ratio ([Unmolded]/[Molded] = [U]/[M]).

Hit #	^1t_R (min)	^2t_R (s)	*RM*	*m/z*	Analyte	MV	[U]/[M]
1	2.49	0.9	99.9	45	Acetic acid	930	2567.50
2	13.57	0.83	98.8	57	1-Hexanol, 2-ethyl-	933	162.01
3	14.89	0.91	96.0	42	Pyrazine, tetramethyl-	882	0.02
4	15.19	1.24	95.6	57	Nonanal	901	44.62
5	1.14	0.58	94.4	44	Carbon dioxide	939	0.03
6	1.37	1.44	93.4	44	Nitrous oxide	942	0.03
7	19.32	0.43	92.8	71	Propanoic acid, 2-methyl-, 3-hydroxy-2,4,4-trimethylpentyl ester	850	26.71
8	12.47	1	92.4	42	Hexanoic acid	927	24.54
9	18.52	0.17	90.8	85	Tridecane	937	20.75
10	19.42	0.06	89.8	71	Hexadecane	928	18.62
11	19.17	0.08	89.5	57	Tridecane, 3-methyl-	866	18.12
12	19.24	0.09	89.1	85	Dodecane, 2,6,10-trimethyl-	932	17.33
13	7.34	1.39	87.9	43	Butanoic acid	852	15.41
14	18.19	0.26	87.3	71	Octane, 2,3,7-trimethyl-	902	14.70
15	18.64	0.16	85.6	57	Tetradecane	868	12.87
16	18.09	0.26	85.0	43	Dodecane, 2-methyl-	889	12.31
17	18.97	0.11	84.4	57	1-Decanol, 2-hexyl-	833	11.83
18	4.52	0.49	84.3	45	Acetoin	802	11.72
19	5.39	0.05	83.8	55	1-Butanol, 3-methyl-	905	0.09
20	19.77	0.03	82.7	57	Nonadecane	876	10.58
21	13.02	0.9	82.6	42	Pyrazine, trimethyl-	918	0.10
22	18.02	0.28	82.5	43	Dodecane, 4-methyl-	879	10.45
23	18.29	0.23	81.9	57	Heptadecane, 2,6,10,14-tetramethyl	871	10.06
24	18.77	0.14	81.1	57	Heptadecane, 2,6-dimethyl-	864	9.58
25	17.02	0.39	78.7	43	Dodecane	898	8.38
26	20.67	0.23	77.9	43	Pentanoic acid, 2,2,4-trimethyl-3-carboxyisopropyl, isobutyl ester	868	7.30
27	18.24	0.31	77.8	57	*E*-2-Octadecadecen-1-ol	847	7.34
28	7.24	0.69	76.0	41	1-Octene	852	3.17
29	17.27	0.41	76.0	106	Undecane, 2,6-dimethyl-	905	6.27
30	16.07	1	74.7	44	Benzene, 1,2-dimethoxy-	872	5.69

Appendix F

This appendix is reproduced from the Electronic Supplementary Material of C. N. Cain, S. Schöneich, R. E. Synovec, Development of an Enhanced Total Ion Current Chromatogram Algorithm to Improve Untargeted Peak Detection, Anal. Chem. 92 (2020), 11365-11373.

Chromatographic Conditions

90-Component test mixture. Separations of a 90-component test mixture (Table F.1) were collected on a GC×GC-TOFMS instrument configured with an Agilent 6890N GC (Agilent Technologies, Palo Alto, CA, USA) and LECO Pegasus III TOFMS (LECO Corporation, St. Joseph, MI, USA). A T-union was used to join the ^{1}D and ^{2}D columns to a pneumatic "pulse valve" (Model 009–0347–900, Parker Hannifin, Hollis, NH, USA), which acts as a total transfer modulator and operates using dynamic pressure gradient modulation (DPGM), a novel form of flow modulation. A Rtx-200 (19 m length, 180 µm d_c, 0.20 µm d_f) was the ^{1}D column and Rxi-1ms (2 m length, 180 µm d_c, 0.18 µm d_f) was the ^{2}D column (Restek, Bellefonte, PA, USA). The TOFMS collected data from 34 *m/z* to 338 *m/z*. The mass acquisition range applied was selected to ensure the capture of signal from all informative ions without interferences, while still maintaining a small data set computationally. The data collection rate was 200 Hz. The electron impact ionization energy was 70 eV and the detector voltage was 1562 V. The test mixture had an original concentration of 10 part-per-thousand (ppth), which was serially diluted in methanol to concentrations of 10 and 1 part-per-million (ppm). The chromatograms collected for the 10 and 1 ppm samples were utilized in the present study. The modulation period (P_M) was 1 s and the pulse width, defined by the length of time that flow is stopped from the pulse valve, was 200 ms. The oven temperature was initially held at 40 ºC for 1 min and ramped to 250 ºC (10 ºC/min rate), where it was held for 1 min. The auxiliary pressure applied to the valve was held constant at 18.0 psig for 1 min and raised to 36.0 psig (0.857 psi/min rate), where it was held for 1 min. The 10 and 1 ppm samples were injected splitless in a 1 µL volume and the inlet flow rate was set to 1 mL/min. Additional experimental detail can be found in a previous report [1].

Yeast cell metabolite extract. Furthermore, a GC×GC-TOFMS separation of yeast cells grown in respiring conditions was collected using an Agilent 6890N GC coupled to a LECO Pegasus III TOFMS with a 4D thermal modulator. Splitless injections in a 1 µL volume were injected and the flow rate was 1 mL/min. The ^{1}D column was a Rtx-5ms (20 m length, 250 µm d_c, 0.5 µm d_f) and the ^{2}D column was a Rtx-200 (2 m length, 180 µm d_c. and 0.2 µm d_f). The ^{1}D column was held at 60 ºC for 0.25 min, increased to 280 ºC at 8 ºC/min, and was held at 280 ºC for 10 min. The ^{2}D column followed the same temperature program except for its initial temperature of 70 ºC. The P_M was 1.5 s and was kept 40 ºC higher than the ^{1}D column. Data was collected at 100 Hz and the TOFMS collected data from 40 *m/z* to 600 *m/z*. More information describing the biological culture, extraction, and derivatization can be found in previous articles [2,3].

Table F.1. List of analytes that made up the 90-component test mixture for the experimentally collected GC×GC-TOFMS chromatograms [1].

Alkanes	**Alkynes**	**Esters**
Hexane	1-hexyne	Ethyl formate
Heptane	1-heptyne	Methyl decanoate
Octane	1-nonyne	Methyl caprylate
Nonane	5-decyne	Methyl salicylate
Decane	**Alcohols**	Ethyl salicylate
Undecane	1-propanol	Methyl laurate
Dodecane	2-butanol	Methyl caproate
Tridecane	1-pentanol	Diethyl phthalate
Tetradecane	2-pentanol	**Ketones**
Pentadecane	1-decanol	2-butanone
Hexadecane	1-tetradecanol	2-pentanone
Pristane	1-octadecanol	2-hexanone
Octadecane	Hexyl alcohol	3-hexanone
Eicosane	2-heptanol	2-heptanone
Halogenated Alkanes	1-octanol	3-heptanone
1,5-dichloropentane	1-nonanol	3-octanone
1-chlorohexane	1-eicosanol	2-decanone
1-bromohexane	Benzyl alcohol	2-undecanone
1-bromoheptane	2-ethyl-1-hexanol	2-dodecanone
1-bromooctane	**Aromatics**	**Aromatics**
1-chlorobutane	Benzene	1,2,4-trimethylbenzene
1,1,1-trichloroethane	Toluene	Anisole
1,2-dichloroethane	3-ethyltoluene	Dibutyl phthalate
Carbon tetrachloride	4-ethyltoluene	a-terpineol (90 %)
Cyclics	Mesitylene	
Methylcyclopentane	Ethylbenzene	
Cyclohexane	Butylbenzene	
Cyclooctane	Isobutylbenzene	
Butylcyclohexane	*t*-butyl benzene	
Bicyclohexyl	Propylbenzene	
2,2,4-trimethylpentane	1-ethylnapthelene	
Alkenes	Bromobenzene	
1-hexene	Cyclohexylbenzene	
Cyclohexene	Diphenylmethane	
Dodecane	p-xylene	
1-undecene	o-xylene	

[1] S. Schöneich, D. V. Gough, T.J. Trinklein, R.E. Synovec, Dynamic pressure gradient modulation for comprehensive two-dimensional gas chromatography with time-of-flight mass spectrometry detection, J. Chromatogr. A. 1620 (2020) 460982. https://doi.org/10.1016/j.chroma.2020.460982.

[2] R.E. Mohler, K.M. Dombek, J.C. Hoggard, E.T. Young, R.E. Synovec, Comprehensive two-dimensional gas chromatography time-of-flight mass spectrometry analysis of metabolites in fermenting and respiring yeast cells, Anal. Chem. 78 (2006) 2700–2709. https://doi.org/10.1021/ac052106o.

[3] R.E. Mohler, K.M. Dombek, J.C. Hoggard, K.M. Pierce, E.T. Young, R.E. Synovec, Comprehensive analysis of yeast metabolite GC×GC-TOFMS data: Combining discovery-mode and deconvolution chemometric software, Analyst. 132 (2007) 756–767. https://doi.org/10.1039/b700061h.

Importing and Preprocessing of Experimentally Collected Chromatograms

Experimentally collected GC×GC-TOFMS chromatograms were imported into Matlab 2019b (Mathworks, Inc., Natick, MA, USA) using an in-house software, which converts the data from the ChromaTOF (LECO) format into Matlab variables [4]. Once chromatograms are unfolded into their vector form, baseline correction is performed on each *m/z* using a rolling ball, which subtracts low frequency noise from the data [4,5]. This technique operates as a "ball" with a fixed radius that rolls along the length of the chromatogram to extract the baseline [4,5]. The appropriate ball size must be chosen to preserve relevant chemical information and avoid overfitting the chromatogram. Optimization of the appropriate "ball radius" for baseline correction of a representative peak is shown in Figure S1. After the chromatograms are baseline corrected, their intensities were centered around zero. The chromatograms were re-folded, and the *m/z* dimension was summed together to create the standard total ion current chromatogram (TIC). A watershed-based algorithm [6,7] was performed for peak detection on the standard TIC using the Matlab Image Processing Toolbox. This approach for peak detection in the standard TIC was chosen because of its widespread use in commercially available software, such as Delta2D [4] and GC Image [6,7].

[4] B. Kehimkar, J.C. Hoggard, L.C. Marney, M.C. Billingsley, C.G. Fraga, T.J. Bruno, R.E. Synovec, Correlation of rocket propulsion fuel properties with chemical composition using comprehensive two-dimensional gas chromatography with time-of-flight mass spectrometry followed by partial least squares regression analysis, J. Chromatogr. A. 1327 (2014) 132–140. https://doi.org/10.1016/j.chroma.2013.12.060.

[5] H.G. Schmarr, J. Bernhardt, Profiling analysis of volatile compounds from fruits using comprehensive two-dimensional gas chromatography and image processing techniques, J. Chromatogr. A. 1217 (2010) 565–574. https://doi.org/10.1016/j.chroma.2009.11.063.

[6] S.E. Reichenbach, M. Ni, V. Kottapalli, A. Visvanathan, Information technologies for comprehensive two-dimensional gas chromatography, Chemom. Intell. Lab. Syst. 71 (2004) 107–120. https://doi.org/10.1016/j.chemolab.2003.12.009.

[7] S.E. Reichenbach, X. Tian, Q. Tao, D.R. Stoll, P.W. Carr, Comprehensive feature analysis for sample classification with comprehensive two-dimensional LC, J. Sep. Sci. 33 (2010) 1365–1374. https://doi.org/10.1002/jssc.200900859.

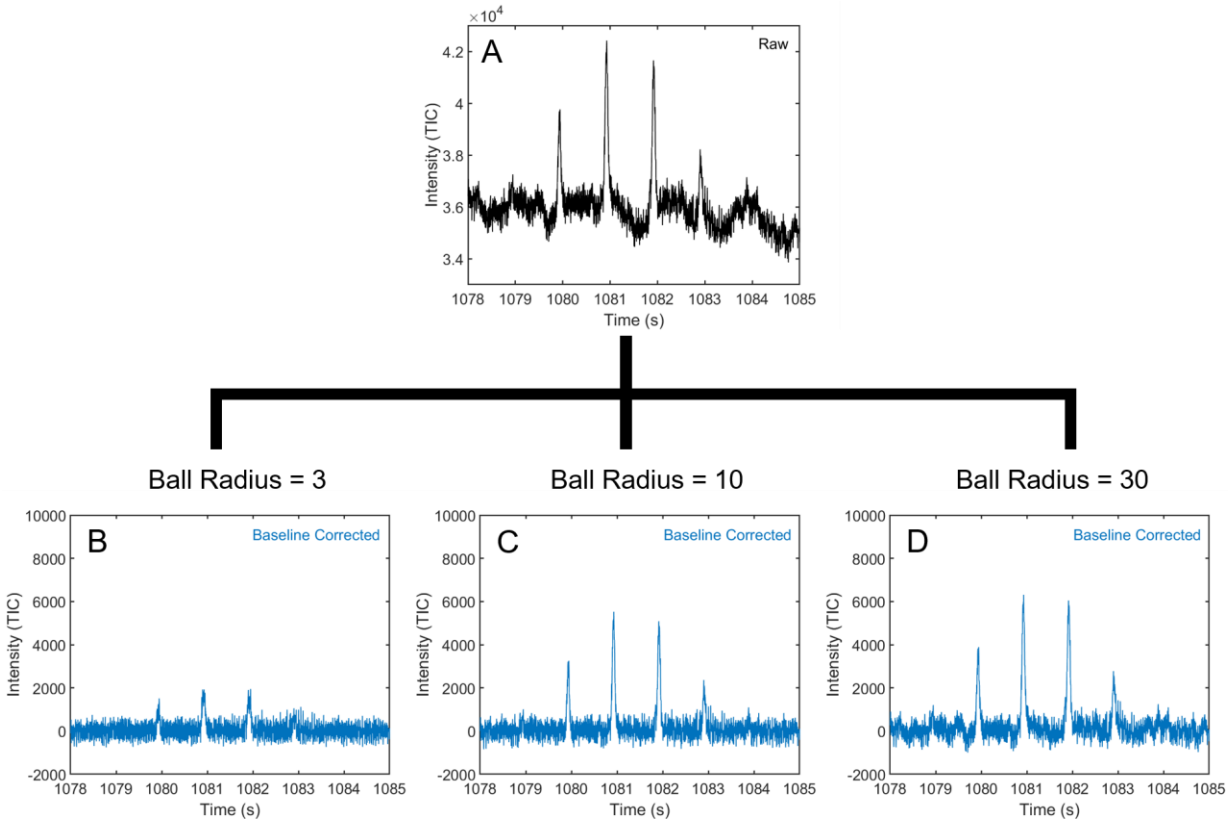


Figure F.1. Schematic illustrating the optimization of the ball radius needed for baseline correction. This optimization process was used for baseline correction of the chromatograms of the 90-component test mixture and yeast cell extract metabolome. (A) The raw chromatogram for eicosane in the 10 ppm test mixture. (B) The chromatogram after applying a ball radius of 3 during baseline correction. This ball radius was determined to be too small since it overfitted the baseline. (C) The chromatogram after applying a ball radius of 10 during baseline correction. The ball radius was determined to be appropriate for baseline correction. (D) The chromatogram after applying a ball radius of 30 during baseline correction. This ball radius was determined to be too large since it did not fully subtract out the low frequency noise in the baseline.

Table F.2. Chromatographic parameters used in GC×GC-TOFMS simulations.

Parameter	Conditions studied
^{1}D separation time, $^1t_{sep}$ (s)	160, 80, 53, 40, 32, 27, 23, 20, 18, 16
Modulation period, P_M (s)	1
Number of components, m	40
Saturation factor, α_{2D}	0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1
Average peak area (TIC)	200
^{1}D peak width-at-base, ^1w_b (s)	4
^{2}D peak width-at-base, ^2w_b (ms)	100
Data collection rate (Hz)	100
S/N relative to the average peak area, S/N_{rel}	0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1, 1.5, 2, 3, 4, 5, 10, 15, 20, 50, 100

Table F.3. List of the analytes selected for the simulation study. Abbreviations: MEOX – methoximation derivatization; TMS – trimethylsilylation derivatization.

Allantoin, 3 TMS	L-Cysteine, 3 TMS	D-(+)-Glucosamine, 6 TMS	L-Isoleucine, TMS	Phenylalanine, 2 TMS
2-Aminoadipic acid, 3 TMS	Cytosine, N,O-2 TMS	D-Glucose, 5 TMS	Isomaltose, TMS	Phosphoric acid, TMS
4-Aminobutanoic acid, 3 TMS	1,3-Diaminopropane, 4 TMS	Glucuronic acid, TMS	Lactic Acid, 2 TMS	L-Proline, 2 TMS
Arabinose, MEOX 4 TMS	Dopamine, 3 TMS	L-Glutamine, 3 TMS	Leucine, 2 TMS	Putrescine, 4 TMS
Aspartic acid, 3 TMS	Erythritol, 4 TMS	Glyceraldehyde, MEOX 2 TMS	Lysine, 3 TMS	Pyruvic acid, MEOX TMS
Benzoic acid, TMS	β-D(-)-Erythrose, TMS	Glyceric acid, 3 TMS	Malic acid, 3 TMS	α-Ribose, TMS
β-Alanine, 2 TMS	Ethanolamine, 3 TMS	Glycine, 3 TMS	Methylcysteine, TMS	Threonine, 3 TMS
Butyric Acid, TMS	Ethylene glycol, 2 TMS	Glycolic acid, 2 TMS	Methylmalonic acid, 2 TMS	Thymine, 2 TMS

Cadaverine, 3 TMS	Ferulic acid, 2 TMS	Glyoxylic acid, MEOX TMS	Norepinephrine, 4 TMS	α-Tocopherol, TMS
Caffeic acid, 3 TMS	D-Fructose, 5 TMS	Hexanoic acid, TMS	L-Norleucine, 2 TMS	Tyramine, 2 TMS
Campesterol, TMS	L-Fucose, 4 TMS	Histidine, N,N,O-3 TMS	Normetanephrine, 3 TMS	L-Tyrosine, 2 TMS
Cholesterol, TMS	Fumaric acid, 2 TMS	4-Hydroxybenzoic acid, 2 TMS	L-Norvaline, 2 TMS	Urea, 2 TMS
Citric acid, 3 TMS	Galactitol, 6 TMS	L-Hydroxyproline, N,O,O-TMS	L-Ornithine, 4 TMS	L-Valine, 2 TMS
Citrulline, TMS	Galactose, TMS	Hydroxypyruvic acid, MEOX 2 TMS	2-Oxobutyric acid, MEOX TMS	Xylitol, 5 TMS
Cystathionine, 4 TMS	D-Gluconic acid, 6 TMS	3-Indoleacetic acid, 2 TMS	L-5-Oxoproline, 2 TMS	Xylose, 4 TMS

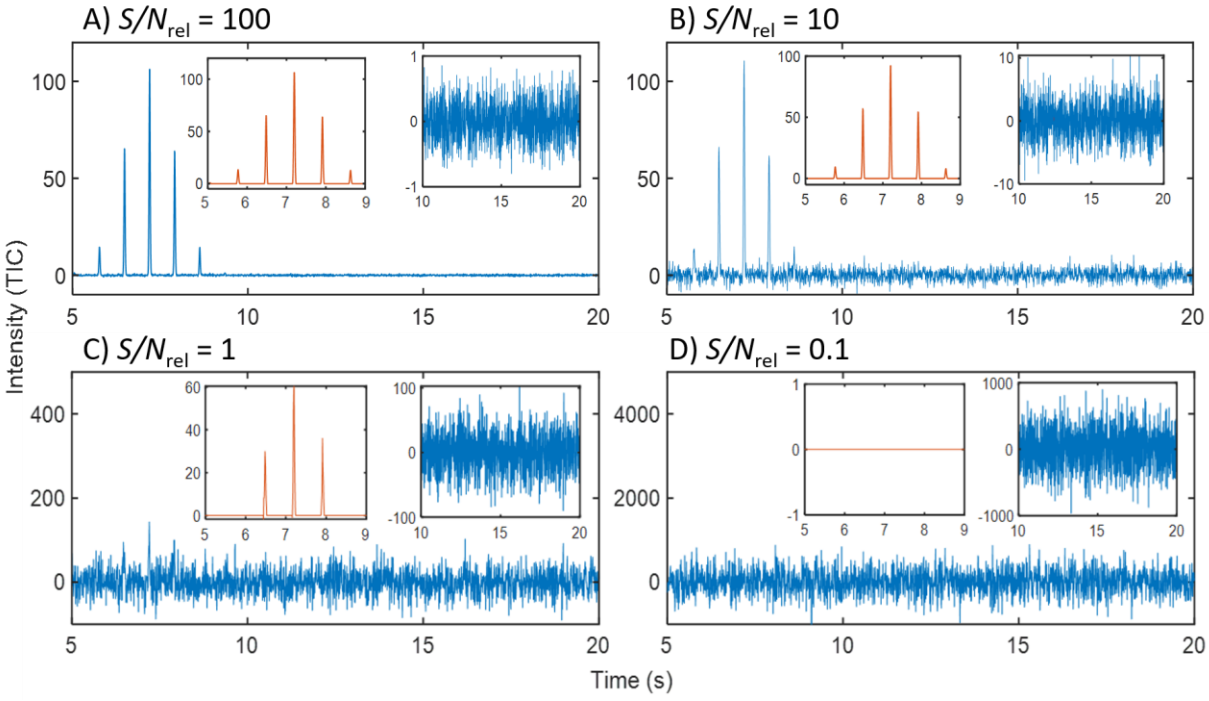


Figure F.2. A simulated unfolded peak with an area of 200 at four different S/N_{rel} values: (A) 100, (B) 10, (C) 1, and (D) 0.1. The left inset in each panel shows the enhanced TIC version of the peak (red) and the right inset zooms in on the baseline noise.

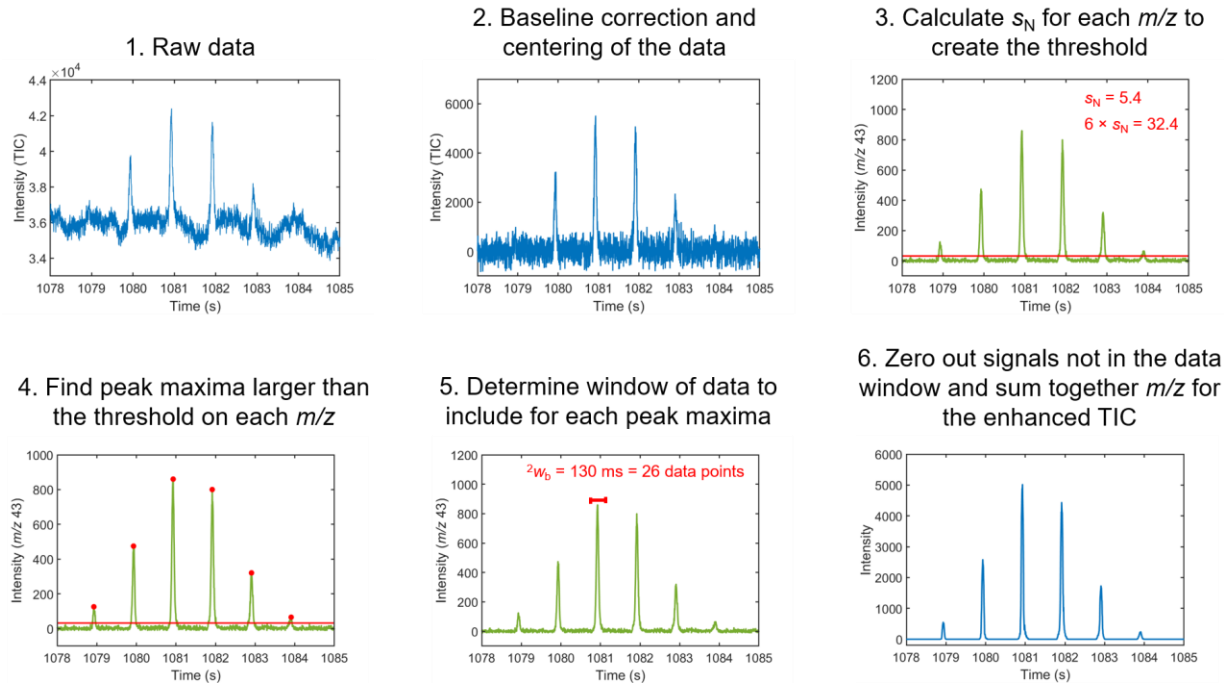


Figure F.3. Schematic illustrating each step of the enhanced TIC algorithm (see Enhanced TIC algorithm) on eicosane in the 10 ppm 90-component test mixture. See Figure F.4 and S5 for more details regarding selection of the appropriate signal threshold in Step 3.

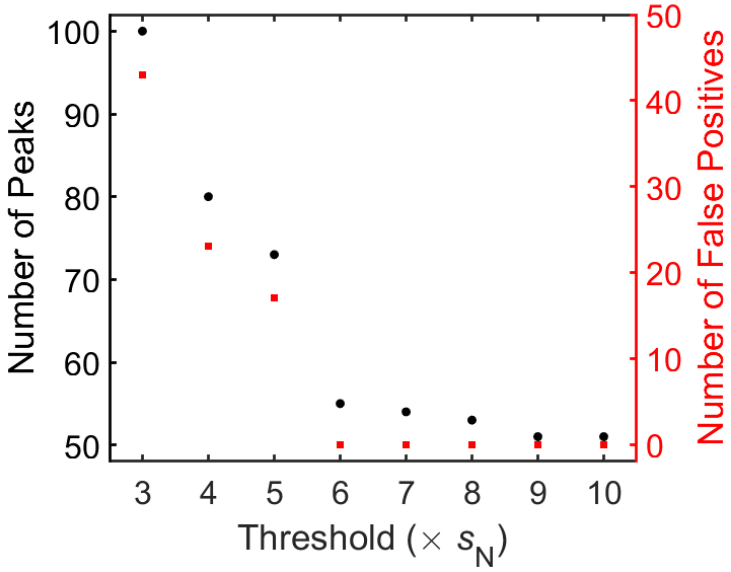


Figure F.4. The number of peaks (black circles) and number of false positives (red squares) as a function of the signal threshold applied during the enhanced TIC method on the 10 ppm test mixture.

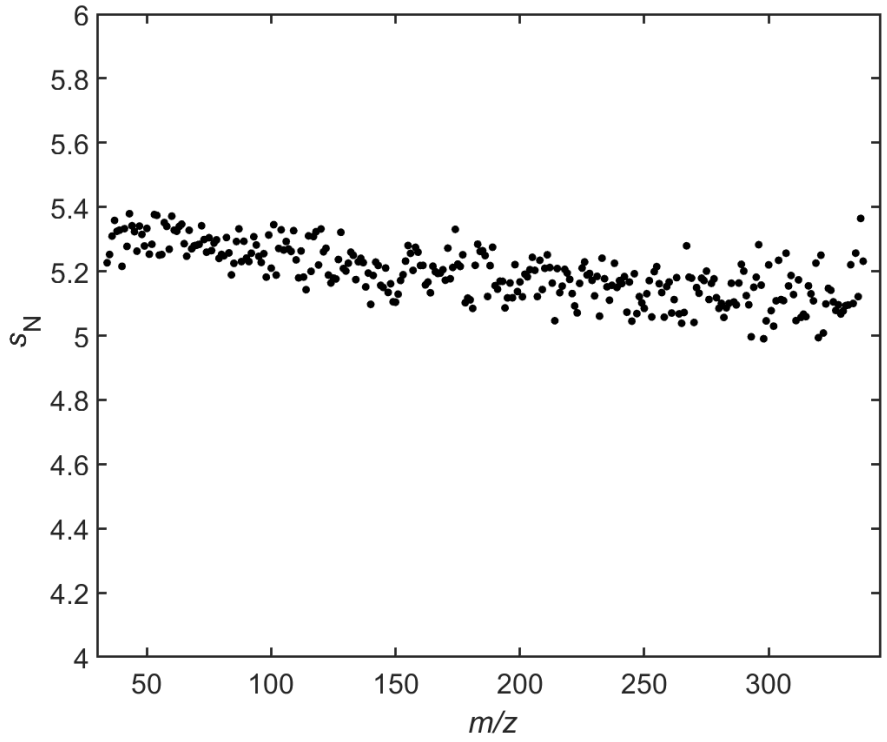


Figure F.5. The standard deviation of the baseline noise (s_N) on each mass channel, *m/z*, in the 10 ppm 90-component test mixture.

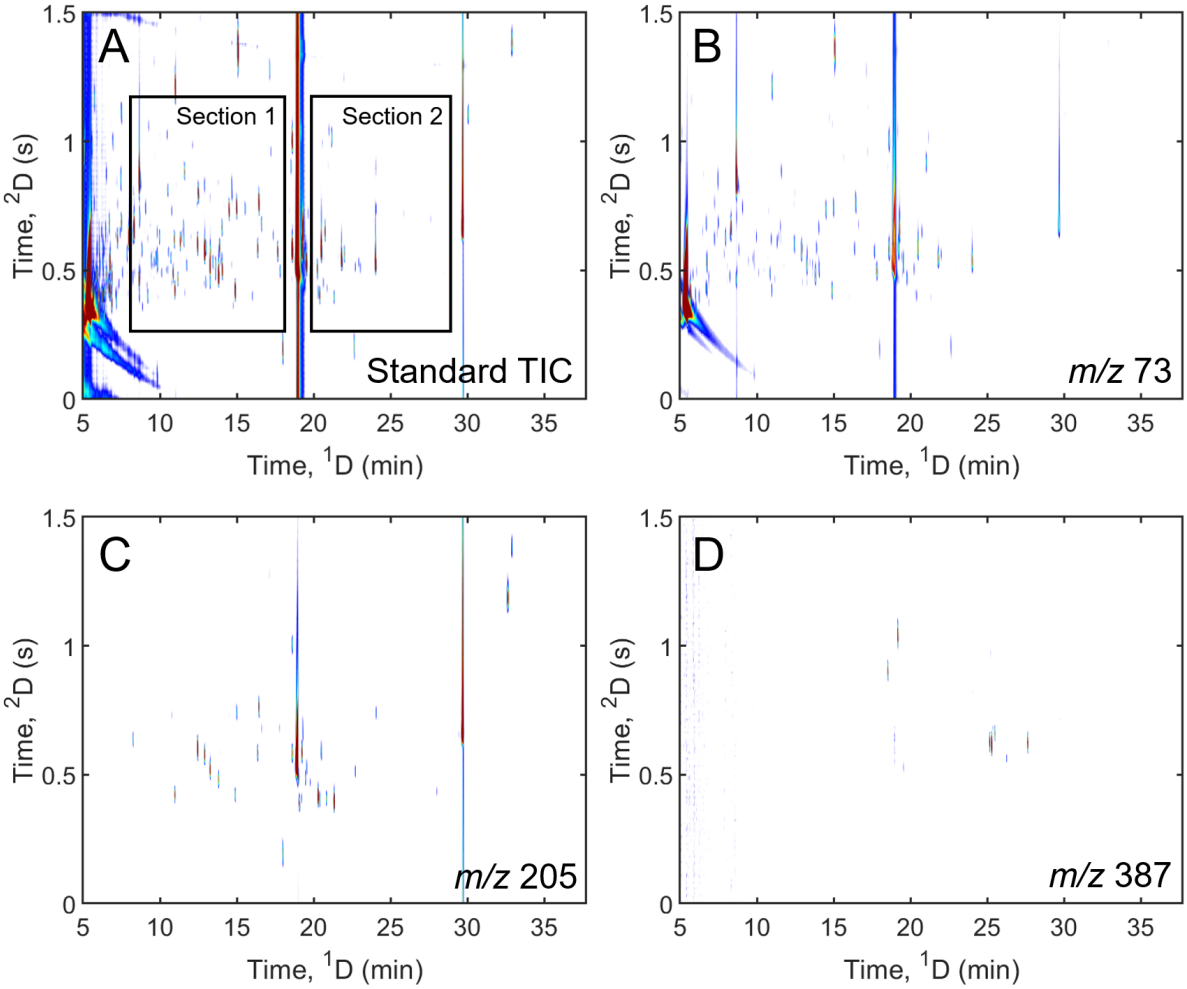


Figure F.6. GC×GC-TOFMS separation of a metabolite extract collected from respiring yeast cells, metabolizing ethanol. (A) The TIC is produced by summing the mass spectral dimension after baseline correction. The two boxes labeled as Section 1 and Section 2 correspond to time windows highlighted in Figure 8.4. (B) The extracted ion current chromatogram (EIC) of the separation in (A) at *m/z* 73. (C) The EIC of the separation in (A) at *m/z* 205, which is selective towards carbohydrates. (D) The EIC of the separation in (A) at *m/z* 387, which is selective towards sugar phosphates.

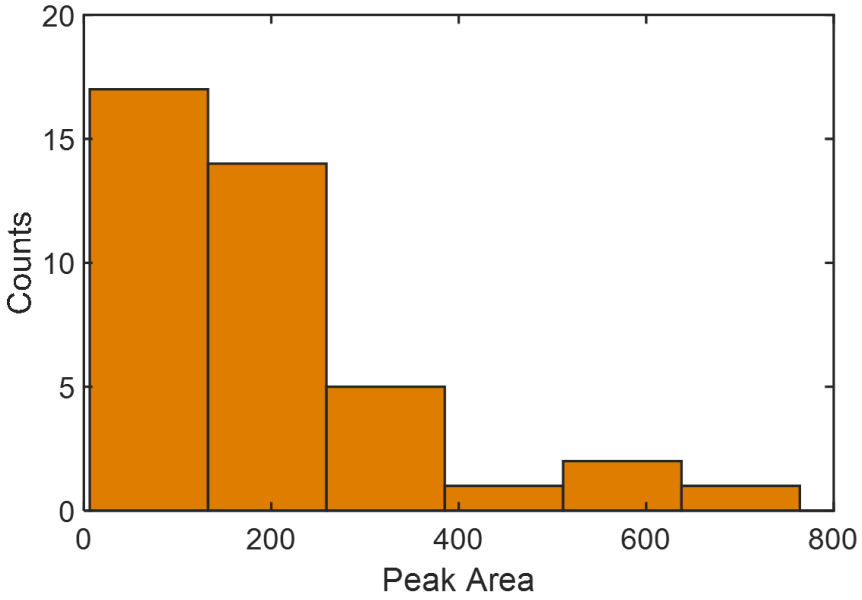


Figure F.7. Exponential distribution of peak areas used for the representative simulations in Figure 8.6.

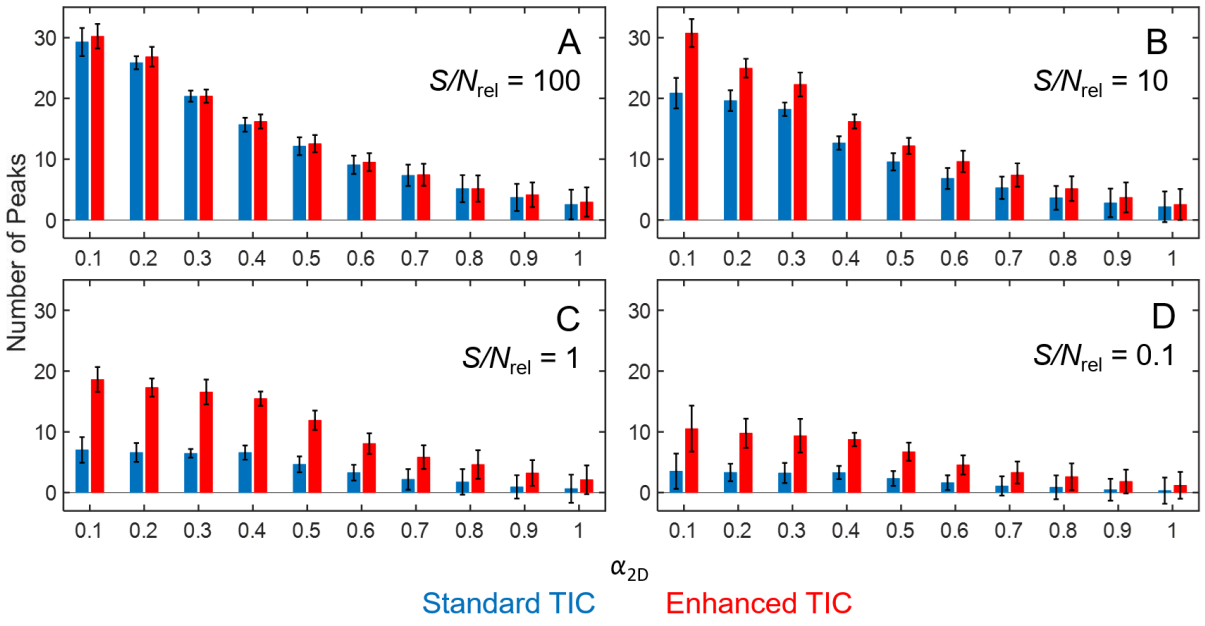


Figure F.8. The number of peaks detected for both the standard TIC (blue) and enhanced TIC (red) as a function of the saturation factor (α_{2D}) simulated at four different S/N_{rel} values: (A) 100, (B) 10, (C) 1, (D) 0.1. Results are shown as the average and standard deviations of 100 simulations for each S/N_{rel} value.

Table F.4. Comparison of the lack of fit (%) between SOT (Eq. 8.7) and the simulated chromatographic results for the standard and enhanced TICs at different S/N_{rel} in Figure 8.7.

S/N_{rel}	Standard TIC	Enhanced TIC
100	5.0	2.5
10	25.2	5.2
1	68.0	20.0
0.1	83.9	53.4

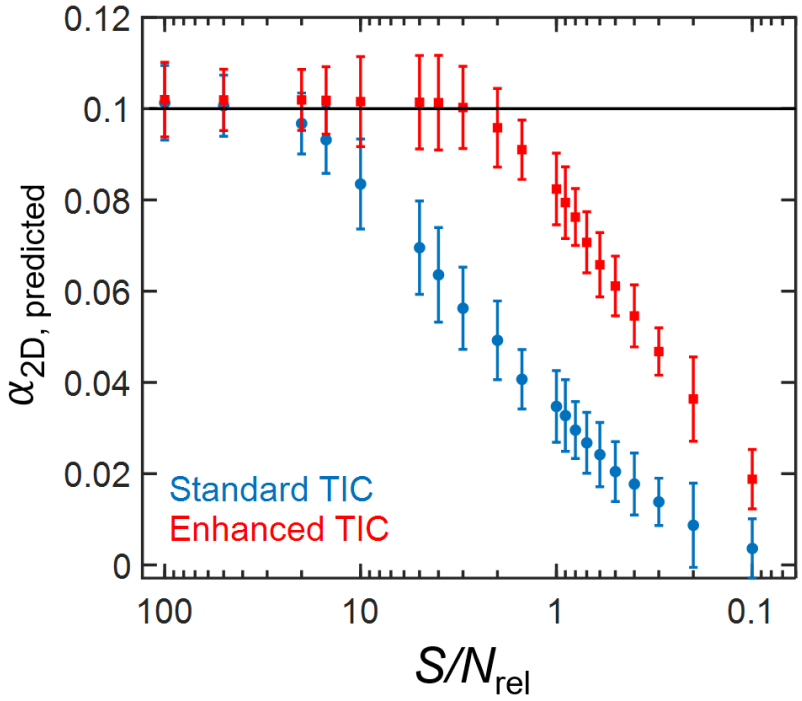


Figure F.9. The effect of S/N_{rel} on the $\alpha_{2D,predicted}$ using the statistical overlap theory for both the standard TIC (blue circles) and enhanced (red squares) TIC. The true α_{2D} for all the simulations was 0.1 (40 analytes in a peak capacity space ($n_{c,2D}$) of 400). Results are shown as the average and standard deviations of 100 simulations.

Appendix G

This appendix is the Supplementary Material for Chapter 9: Enhancing Gas Chromatography-Mass Spectrometry Resolution and Pure Analyte Discovery using Intra-Chromatogram Elution Profile Matching.

Chromatographic Conditions

73-Component test mixture. The chromatographic conditions for this separation can be found in our previous report [1]. Briefly, separations of a 73-component test mixture (Table G.1) were collected on a GC-TOFMS instrument configured with an Agilent 6890N GC (Agilent Technologies, Palo Alto, CA, USA) and LECO Pegasus III TOFMS (LECO Corporation, St. Joseph, MI, USA). An RTX-5 column (30 m × 250 μm × 0.5 μm; Restek, Bellefonte, PA, USA) was used for the separation. The carrier gas was ultra-high purity helium (Grade 5, 99.999 %; Praxair, Seattle, WA, USA) and operated at a constant flow rate of 2 mL/min. The temperature program started with a 35 ºC hold for 1 min before ramping to 280 ºC at 33 ºC/min, where the temperature was held again for 1 min. A 0.1 μL aliquot of the diluted 73-component test mixture (1:100 in acetone) was injected into the inlet at a split ratio of 200:1. The inlet temperature was 280 ºC. After a 100 s solvent delay, the TOFMS collected *m/z* 40 – 230 at 100 Hz. The ion source temperature was 250 ºC and the electron impact voltage was 70 eV.

115-Component test mixture. The chromatographic conditions for this separation can be found in our previous report [2]. Separations of a 115-component test mixture (Table G.2) were collected on an LTM-GC-TOFMS instrument configured with an Agilent LTM column module, Agilent 6890N GC, and LECO Pegasus III TOFMS. An Agilent DB-5 column (10 m × 100 μm × 0.4 μm) was used for the separation. The carrier gas was ultra-high purity helium (Grade 5, 99.999 %) and operated at a constant flow rate of 2 mL/min. The temperature program started with a 50 ºC hold for 30 s before ramping to 300 ºC at 250 ºC/min, where the temperature was held again for 1 min. A 0.05 μL aliquot of the neat test mixture was injected into the inlet at a split ratio of 300:1. The inlet temperature was 300 ºC. The TOFMS collected *m/z* 33 – 250 at 500 Hz. The ion source temperature was 250 ºC and the electron impact voltage was 70 eV.

Aerospace fuel sample. Separations of an aerospace fuel were collected on a GC-TOFMS instrument configured with Agilent 7890 GC and LECO Pegasus BT 4D TOFMS. A Restek Rxi-5Sil MS column (60 m × 250 μm × 0.25 μm) was used for the separation. The carrier gas was ultra-high purity helium (Grade 5, 99.999 %) and operated at a constant flow rate of 2 mL/min. The temperature program started with a 40 ºC hold for 2 min before ramping to 50 ºC at 5 ºC/min, where the temperature was held again for 5 min. Then, a second temperature ramp (10 ºC/min) was employed to reach a target temperature of 200 ºC. The temperature of the GC oven remained constant at 200 ºC for the last 15 mins. A 0.01 μL aliquot of the fuel was injected into the inlet at a split ratio of 150:1. The inlet temperature was 250 ºC. After a 10 s solvent delay, the TOFMS collected *m/z* 45 – 300 at 500 Hz. The ion source temperature was 225 ºC and the transfer line temperature was 285 ºC. The electron impact voltage was set to 70 eV.

Table G.1. List of analytes that made up the 73-component test mixture along with their boiling point (BP) and vendor [1].

Compound Name	BP (°C)	Vendor
pentane	36	J.T. Baker
ethyl formate	54	Sigma-Aldrich
acetone	56	Sigma-Aldrich
chloroform	61	Fischer
hexane	69	Sigma-Aldrich
methylcyclopentane	72	Fluka
2-butanone	80	Sigma-Aldrich
benzene	80	Fischer
cyclohexane	81	Sigma-Aldrich
1-propanol	97	Sigma-Aldrich
heptane	98	Fischer
methylcyclohexane	101	Sigma-Aldrich
2-pentanone	102	Sigma-Aldrich
toluene	111	Sigma-Aldrich
2-butanol	117	Sigma-Aldrich
octane	126	Sigma-Aldrich
3-hexanone	128	Sigma-Aldrich
chlorobenzene	132	Alfa-Aesar
1-chlorohexane	135	Sigma-Aldrich
ethylbenzene	136	Sigma-Aldrich
1-pentanol	137	Sigma-Aldrich
xylenes (*o-, m-,* and *p-*)	139	Mallinckrodt
3-heptanone	141	Sigma-Aldrich
cyclooctane	150	Sigma-Aldrich
nonane	151	Sigma-Aldrich
propylbenzene	152	Sigma-Aldrich
1-bromohexane	156	Sigma-Aldrich
bromobenzene	156	Sigma-Aldrich
1-hexanol	158	Sigma-Aldrich
2-heptanol	160	Fluka
mesitylene	163	Sigma-Aldrich
3-octanone	167	Sigma-Aldrich
tert-butyl benzene	167	Sigma-Aldrich
methyl hexanoate	168	Sigma-Aldrich
isobutylbenzene	170	Sigma-Aldrich
sec-butyl benzene	173	Sigma-Aldrich
decane	174	Sigma-Aldrich
1-bromoheptane	179	Sigma-Aldrich
butylcyclohexane	181	Sigma-Aldrich
butylbenzene	183	Sigma-Aldrich
methyl caprylate	193	Sigma-Aldrich
1-octanol	194	Sigma-Aldrich
2-nonanone	194	Sigma-Aldrich
undecane	196	Sigma-Aldrich
1-bromooctane	199	Sigma-Aldrich
1,3,5-trichlorobenzene	208	Sigma-Aldrich
2-decanone	209	Sigma-Aldrich
1-nonanol	215	Sigma-Aldrich
dodecane	216	Sigma-Aldrich

naphthalene	218	Sigma-Aldrich
1,2,3-trichlorobenzene	219	Sigma-Aldrich
methyl salicylate	223	Alfa-Aesar
methyl decanoate	224	Eastman Chemicals
ethyl salicylate	227	Alfa-Aesar
bicyclohexyl	227	Sigma-Aldrich
1-decanol	229	Sigma-Aldrich
1-geraniol	230	Sigma-Aldrich
2-undecanone	231	Sigma-Aldrich
tridecane	235	Sigma-Aldrich
cyclohexylbenzene	239	Sigma-Aldrich
2-dodecanone	245	Sigma-Aldrich
tetradecane	254	Fluka
1-dodecanol	256	Acros Organics
methyl laurate	267	Sigma-Aldrich
pentadecane	271	Alfa-Aesar
hexadecane	287	Sigma-Aldrich
1-tetradecanol	289	Sigma-Aldrich
2-pentadecanone	294	Sigma-Aldrich
pristane	296	Sigma-Aldrich
1-hexadecanol	344	Sigma-Aldrich
adamantane	sublimes	Sigma-Aldrich

Table G.2. List of analytes that made up the 115-component test mixture along with their boiling point (BP) and vendor [2].

Compound Name	BP (°C)	Vendor
pentane	36	J.T. Baker
cyclopentane	50	Sigma-Aldrich
ethyl formate	54	Sigma-Aldrich
2-methylpentane	60	Aldrich
1-chloroform	61	Fischer
1-hexene	63	Sigma-Aldrich
1-bromoheptane	68	Sigma-Aldrich
hexane	69	Sigma-Aldrich
1-hexyne	71	Aldrich
methylcyclopentane	72	Fluka
1,1,1-trichloroethane	75	Aldrich
carbon tetrachloride	77	Aldrich
1-chlorobutane	78	Aldrich
2-butanone	80	Sigma-Aldrich
benzene	80	Fischer
1-bromooctane	81	Sigma-Aldrich
cyclohexane	81	Sigma-Aldrich
2-methyl-2-propanol	82	Aldrich
1,2-dichloroethane	83	Aldrich
isopropyl alcohol	83	EMD
1,6-dichlorohexane	90	Aldrich
1-heptene	94	Aldrich
1-propanol	97	Sigma-Aldrich
heptane	98	Sigma-Aldrich
2,2,4-trimethylpentane	99	Sigma-Aldrich

methylcyclohexane	101	Sigma-Aldrich
tert-amyl alcohol	102	Aldrich
2-pentanone	102	Sigma-Aldrich
isobutyl alcohol	108	Baker
1-heptyne	109	Aldrich
toluene	111	Sigma-Aldrich
neopentyl alcohol	114	Aldrich
2,3,4-trimethylpentane	114	Aldrich
2-butanol	117	Sigma-Aldrich
1-butanol	117	Aldrich
2-pentanol	119	Aldrich
octane	126	Sigma-Aldrich
3-hexanone	128	Sigma-Aldrich
2-hexanone	128	Aldrich
m-xylene	128	Aldrich
cis-1,2-dimethylcyclohexane	130	Sigma-Aldrich
chlorobenzene	131	Alfa-Aesar
1-chlorohexane	134	Sigma-Aldrich
ethylbenzene	136	Sigma-Aldrich
1-pentanol	137	Sigma-Aldrich
3-heptanone	141	Sigma-Aldrich
o-xylene	145	Aldrich
cyclooctane	150	Sigma-Aldrich
nonane	151	Sigma-Aldrich
1-nonyne	151	Aldrich
methyl caproate	151	Aldrich
2-heptanone	152	Aldrich
anisole	154	Aldrich
bromobenzene	155	Sigma-Aldrich
1-bromohexane	158	Sigma-Aldrich
1-hexanol	158	Sigma-Aldrich
propylbenzene	159	Sigma-Aldrich
2-heptanol	160	Fluka
R-(-)-2,6-dimethyloctane	160	Aldrich
cyclohexanol	161	Aldrich
mesitylene	163	Sigma-Aldrich
3-octanone	167	Sigma-Aldrich
tert-butyl benzene	167	Sigma-Aldrich
1,2,4-trimethylbenzene	169	Aldrich
isobutylbenzene	170	Sigma-Aldrich
sec-butyl benzene	173	Sigma-Aldrich
decane	174	Sigma-Aldrich
5-decyne	178	Aldrich
butylcyclohexane	181	Sigma-Aldrich
butylbenzene	183	Sigma-Aldrich
1,2-propanediol	188	Aldrich
p-xylene	189	Aldrich
methyl caprylate	193	Sigma-Aldrich
1-undecene	194	Fluka
1-octanol	194	Sigma-Aldrich
2-nonanone	194	Sigma-Aldrich
undecane	196	Sigma-Aldrich
benzyl alcohol	205	Aldrich

1,3,5-trichlorobenzene	208	Sigma-Aldrich
2-decanone	209	Sigma-Aldrich
1-octadecanol	210	Aldrich
dodecene	214	Sigma
1-nonanol	215	Sigma-Aldrich
dodecane	216	Sigma-Aldrich
naphthalene	218	Sigma-Aldrich
1,2,3-trichlorobenzene	218	Sigma-Aldrich
eicosane	220	Aldrich
methyl salicylate	220	Alfa-Aesar
methyl decanoate	224	Eastman Chemicals
bicyclohexyl	227	Sigma-Aldrich
1-decanol	229	Sigma-Aldrich
1-geraniol	230	Sigma-Aldrich
2-undecanone	231	Sigma-Aldrich
ethyl salicylate	234	Alfa-Aesar
tridecane	235	Fluka
cyclohexylbenzene	239	Sigma-Aldrich
2-dodecanone	245	Sigma-Aldrich
1,2,4,5-tetrachlorobenzene	246	Aldrich
tetradecane	254	Fluka
1-dodecanol	256	Acros Organics
methyl laurate	262	Sigma-Aldrich
pentadecane	271	Alfa-Aesar
hexadecane	287	Sigma-Aldrich
1-tetradecanol	289	Sigma-Aldrich
2-pentadecanone	293	Sigma-Aldrich
pristane	296	Sigma-Aldrich
diethyl phthalate	299	Aldrich
heptadecane	302	ICN Biomedicals
benzophenone	305	Sigma-Aldrich
octadecane	317	Fluka
nonadecane	330	Acros
phenanthrene	340	Eastman
1-hexadecanol	344	Sigma-Aldrich
1-eicosanol	372	Aldrich
adamantane	sublimes	Sigma-Aldrich

[1] B.D. Fitz, B.C. Reaser, D.K. Pinkerton, J.C. Hoggard, K.J. Skogerboe, R.E. Synovec, Enhancing gas chromatography-time of flight mass spectrometry data analysis using two-dimensional mass channel cluster plots, Anal. Chem. 86 (2014) 3973–3979. https://doi.org/10.1021/ac5004344.

[2] B.D. Fitz, R.E. Synovec, Extension of the two-dimensional mass channel cluster plot method to fast separations utilizing low thermal mass gas chromatography with time-of-flight mass spectrometry, Anal. Chim. Acta. 913 (2016) 160–170. https://doi.org/10.1016/j.aca.2016.01.045.

Table G.3. List of the 45 analytes selected for the simulation study.

pentane	1,6-dichlorohexane	2-hexanone
hexane	methylcyclopentane	2-heptanone
octane	cyclohexane	*m*-xylene
nonane	butylcyclohexane	2,3,6,7-tetramethyloctane
decane	*cis*-1,2-dimethylcyclohexane	2-octene-4-ol
undecane	2,2,4-trimethylpentane	3,4,5-trimethylheptane
dodecane	2,3,4-trimethylpentane	3,4-diethylhexane
tridecane	2-methylpentane	4,5-dipropyloctane
tetradecane	1-octanol	4-ethyl-2,2,6,6-tetramethylheptane
pentadecane	1-hexadecanol	4-propylheptane
pristane	neopentyl alcohol	butanoic acid
octadecane	1-undecene	docosane
eicosane	3-hexanone	heptanoic acid
heptadecane	2-nonanone	isopentyl alcohol
nonadecane	2-pentadecanone	pentacosane